



INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ⁶ : C07K 1/04		A1	(11) International Publication Number: WO 97/11958 (43) International Publication Date: 3 April 1997 (03.04.97)
(21) International Application Number: PCT/US96/15516		(74) Agents: LEWIS, Donald, G. et al.; The Scripps Research Institute, 10550 North Torrey Pines Road, TPC-8, La Jolla, CA 92037 (US).	
(22) International Filing Date: 27 September 1996 (27.09.96)		(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, TJ, TM, TR, TT, UA, UG, US, UZ, VN, ARIPO patent (KE, LS, MW, SD, SZ, UG), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).	
(30) Priority Data: 60/004,563 29 September 1995 (29.09.95) US		(60) Parent Application or Grant (63) Related by Continuation US 60/004,563 (CIP) Filed on 29 September 1995 (29.09.95)	
(71) Applicant (for all designated States except US): THE SCRIPPS RESEARCH INSTITUTE [US/US]; 10550 North Torrey Pines Road, La Jolla, CA 92037 (US).		Published With international search report. Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.	
(72) Inventors; and (75) Inventors/Applicants (for US only): KENT, Stephen, B., H. [NZ/US]; 6046 La Jolla Hermosa Avenue, La Jolla, CA 92037 (US). MUIR, Tom, W. [GB/US]; 1153 Wilbur Avenue, San Diego, CA 92109 (US). DAWSON, Philip, E. [US/US]; 8560C Via Mallorca Drive, La Jolla, CA 92037 (US). FITZGERALD, Michael, C. [US/US]; 5360 Toscana Way, G-116, San Diego, CA 92122 (US).			

(54) Title: PROTEIN SIGNATURE ANALYSIS

(57) Abstract

A protein signature analysis is obtained using a peptide ladder library. The molecular signature of a protein is defined to be that subsequence of amino acid positions within the protein which are essential for the protein to bind to a target molecule. The molecular signature may be determined by screening a peptide ladder library which corresponds to the protein against the target molecule. The peptide ladder library is a library of m peptides wherein each peptide has an amino acid sequence of length m corresponding to an amino acid sequence of the protein, with one exception, viz. peptide_m has a substitute amino acid at position_m and the substitute amino acid is attached by a labile bond to its neighboring amino acid. Screening the peptide ladder library against the target molecule results in a division of the original mixture into a positive (functional) pool and a negative (non-functional) pool. The pools are separated and subjected to cleavage to obtain cleavage products. Analysis of the cleavage products by mass spectrometry identifies the positions that are essential for the protein to bind to its molecular target.

Parent peptide sequence: PFKK-[**GDILRIRDKP**]-EEAc_pRLKLKAR

Target Peptide Array

PFKK **[X]** I L R I R D K P EEAc_pRLKLKAR
 PFKK G **[X]** I L R I R D K P EEAc_pRLKLKAR
 PFKK G D **[X]** I R I R D K P EEAc_pRLKLKAR
 PFKK G D I **[X]** I R D K P EEAc_pRLKLKAR
 PFKK G D I L **[X]** R D K P EEAc_pRLKLKAR
 PFKK G D I L R **[X]** D K P EEAc_pRLKLKAR
 PFKK G D I L R I **[X]** K P EEAc_pRLKLKAR
 PFKK G D I L R I R **[X]** P EEAc_pRLKLKAR
 PFKK G D I L R I R D **[X]** EEAc_pRLKLKAR

[X] =Gly-S β Ala

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgyzstan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic of Korea	SD	Sudan
CF	Central African Republic	KR	Republic of Korea	SE	Sweden
CG	Congo	KZ	Kazakhstan	SG	Singapore
CH	Switzerland	LI	Liechtenstein	SI	Slovenia
CI	Côte d'Ivoire	LK	Sri Lanka	SK	Slovakia
CM	Cameroon	LR	Liberia	SN	Senegal
CN	China	LT	Lithuania	SZ	Swaziland
CS	Czechoslovakia	LU	Luxembourg	TD	Chad
CZ	Czech Republic	LV	Latvia	TG	Togo
DE	Germany	MC	Monaco	TJ	Tajikistan
DK	Denmark	MD	Republic of Moldova	TT	Trinidad and Tobago
EE	Estonia	MG	Madagascar	UA	Ukraine
ES	Spain	ML	Mali	UG	Uganda
FI	Finland	MN	Mongolia	US	United States of America
FR	France	MR	Mauritania	UZ	Uzbekistan
GA	Gabon			VN	Viet Nam

- 1 -

PROTEIN SIGNATURE ANALYSIS

Description

5

Technical Field of the Invention:

The present invention relates to methods for analyzing, altering, and controlling the structural basis for protein binding to target molecules. More 10 particularly, the present invention is directed to peptide ladder libraries corresponding to a protein, protein fragment, or other bioactive peptide and to the use of peptide ladder libraries for obtaining a protein signature analysis.

15

Government Rights:

This invention was made with government support under Grants No. PO1GM 48870, HL31950, and RO1GM 48897 awarded by the National Institutes of Health. 20 The U.S. government has certain rights in the invention.

Background of the Invention:

One of the major strategies for determining the 25 relationship between the chemical structure of a peptide and its biological activity is to systematically alter the covalent structure and observe the effect on function. Through the use of chemical synthesis, a wide variety of modifications 30 can be made. For example, ^oN-methylation and the use of ester bonds can probe backbone interactions (Arad et al. *Biopolymers* 1990, 29, 1633-1649; Bramson et

- 2 -

al. *J. Biol. Chem.* **1985**, *260*, 15452-15457; Caporale et al. In: *Peptides: Structure and Function, Proceedings of the Tenth American Peptide Symposium*; Marshall, G.F. Ed. Escom: Leiden: The Netherlands, 1988, pp. 449-451), while sidechain contributions can be probed using D-amino acid or Alanine/Glycine substitutions (Konishi et al. In: *Peptides: Structure and Function, Proceedings of the Tenth American Peptide Symposium*, Marshall, G.F. Ed. Escom, Leiden: The Netherlands, 1988, pp. 479-481; Tam et al. In *Peptides: Proceedings of the Eleventh American Peptide Symposium*; Rivier, J. E.; Marshall, G. R. Ed.; Escom: Leiden, The Netherlands, 1990. pp 75-77). As traditionally practiced, a separate analogue must be prepared and assayed for each position in the peptide sequence that is to be studied.

An alternative, currently popular method of studying peptides is through combinatorial chemistry. This approach has had a major impact on the study of the molecular basis of peptide activity and has contributed to the search for new biologically active peptides (Thompson et al. *Chem. Rev.* **1996**, *96*, 555-600; Gordon et al. *J. Med. Chem.* **1994**, *37*, 1385-1401; Scott et al. *Curr. Op. Biotech* **1994**, *5*, 40-48) 'Multiple Peptide Synthesis' has extended the traditional approach by allowing peptides to be synthesized simultaneously (Geysen et al. *J. Proc. Natl. Acad. Sci. USA* **1984**, *81*, 3998-4001; Houghten et al. *Proc Natl. Acad. Sci. USA* **1985**, *82*, 5131-5134). The individual peptide products are spatially separated and can be analyzed either attached to a

- 3 -

solid support or in solution. Established 'split synthesis' (Furka et al. *Int. J. Pept. Prot. Res.* 1991, 37, 487-494; Lam et al. *Nature* 1991, 354, 82-84) procedures allow for the rapid generation of huge numbers of peptide sequences through the repetition of a simple divide, couple and recombine process. The compositional diversity made possible by this approach is advantageous for the discovery of new 'lead' compounds since, in principle, all possible structural variants can be explored for the desired activity and only the few active oligomers of interest need to be individually identified (Furka et al. *Int. J. Pept. Prot. Res.* 1991, 37, 487-494; Lam et al. *Nature* 1991, 354, 82-84). However, where information about a complete set of functional and non-functional components is desired over many positions in a peptide sequence, such libraries are too complex to fully characterize and may have limited utility.

A more systematic investigation of the molecular basis of peptide function requires a different type of molecular diversity. Instead of a peptide mixture of high compositional diversity, it would be useful to construct an array of peptides which differ from each other in a precise and defined manner. In principle, one way to access this population would be as a minor fraction of a large, fully combinatorial library. For example, such an array of analogues could consist of all peptides which differ from a target sequence by a single amino acid substitution at each position in a peptide sequence (cf. 'Ala scans'). By removing this defined subset of

- 4 -

analogues from the context of a complex, fully
combinatorial mixture of peptides, handling and
analysis would be greatly simplified and a more
useful profile of the effects of substituting the
5 amino acid throughout the peptide chain would be
obtained. Current split resin methods do not allow
for this type of control over the composition of a
peptide library. (Furka et al. *Int. J. Pept. Prot.*
Res. 1991, 37, 487-494; Lam et al. *Nature* 1991, 354,
10 82-84).

Typically, to investigate the molecular basis of
protein function systematic modifications are made to
the protein structure and the effects of those
modifications on the properties of the protein are
evaluated. Site-directed mutagenesis (Smith et al.
15 *Angew. Chem. Int. Ed. Engl.* 1994, 33, 1214-1220) has
been the principle tool used to implement this
approach and has given many insights into the
contribution of individual sidechains to protein
function. In particular, 'alanine scanning' (Wells
et al. *Methods in Enzymology* 1991, 202, 390-411) has
been used to identify specific amino acid sidechains
involved in ligand binding interactions. This
technique involves the sequential substitution of
20 native amino acids by individual alanine residues
which are regarded as functionally and structurally
neutral. To extend the repertoire of modifications
beyond the twenty genetically encoded amino acids,
methods have been developed to substitute non-natural
25 groups into proteins (Noren et al. *Science*. 1989.
30 244, 182-185). Although a variety of both novel

- 5 -

sidechain and backbone modified proteins have been generated, there are apparent limits to the modifications possible using the methods of molecular biology and ribosomal synthesis (Ellman et al.

5 *Science* 1991, 255, 197-200; Cornish et al. *Angew Chem Int. Ed. Engl.* 1995, 34, 621-633).

Recent advances in the total synthesis of polypeptides have opened the world of proteins to direct application of the tools of organic chemistry
10 (Schnölzer et al. *Science* 1992, 256, 221-225; Jackson et al. *Science* 1994, 266, 243-247; Dawson et al *Science* 1994, 266, 776-779; Canne et al. *J. Am. Chem. Soc.* 1995, 117, 2998-3007; Liu et al *J. Am. Chem. Soc.* 1995, 118, 307-312; Englebretsen et al. *Tet. Lett.* 1995, 36, 8871-8874). Using total chemical synthesis, a variety of protein analogues has been synthesized. Of particular note have been proteins containing β -turn mimics (Baca et al. *Prot. Sci.*

15 *1993*, 2, 1085-1091), N-methylated amino acids (Rajarathnam et al. *Science* 1994, 264, 90-92), modified backbone atoms (Baca et al. *J. Am. Chem. Soc.* 1995, 117, 1881-1887), and mirror image proteins composed entirely of D-amino acids (Zawadzke et al.

20 *J. Am. Chem. Soc.* 1992, 114, 4002-4003; Milton et al. *Science* 1992, 256, 1445-1448; Fitzgerald et al. *J. Am. Chem. Soc.* 1995, 117, 11075-11080; Schumaacher et al. *Science* 1996, 271, 1854-1857). In addition,

25 important insights into the mechanism of action of enzymes have been attained through the total chemical synthesis of unique analogues (Baca et al. *Proc. Natl. Acad. Sci. U.S.A.* 1993, 90, 11638-11642).

- 6 -

Although structure-function relationships in proteins can be studied using individual analogues prepared by either recombinant or chemical techniques, development of a profile of effects across the whole protein molecule is hindered by the time and effort required to generate and analyze multiple protein analogues (Matthews et al. *Ann. Rev. Biochem.* 1993, 62, 139-160). The use of combinatorial oligonucleotide synthesis in conjunction with protein expression in bacteria (Reidhaar-Olsen et al. *Science* 1988, 241, 53-57; Gregoret et al. *Proc. Natl. Acad. Sci. USA.* 1993, 90, 4246-4250) or on phage (Scott et al. *Science* 1990, 249, 386-390; Lowman, H. B. Bass, S.H.; Simpson, N.; Wells, J. A. *Biochemistry* 1991, 30 10832-10838) has provided a powerful method for studying large numbers of analogue proteins. These techniques allow pools of expressed proteins to be probed for a desired function. With appropriate screening procedures, a statistical sampling of numerous functional protein variants can be analyzed and identified (Gu et al. *Protein Science* 1995, 4, 1108-1117). This strategy has proved to be powerful for generating variant proteins with new or optimized functions (Lowman et al. *J. Mol. Biol.* 1993, 234, 564-578; Rebar et al. *Science* 1994, 263, 671-673). However, studies designed to elucidate the molecular basis of protein function have been complicated by the necessarily incomplete characterization of the numerous protein analogues generated, and also by limitation to the naturally encoded amino acids.

In applying molecular diversity to the study

- 7 -

protein function it would be useful to combine the valuable information gained by systematic modification through chemical synthesis with the advantages of combinatorial methods.

5 What is needed is an integrated approach to the preparation of a defined array of peptide and protein analogues in a single synthesis, their functional separation into active and inactive pools, and a simple one step readout of the composition of the
10 self-encoded mixtures.

Summary of the Invention:

There are three aspects to the invention:

1. A combinatorial method for synthesizing a peptide ladder library corresponding to a
15 protein, protein fragment, or other bioactive peptide.
2. A method for screening the peptide ladder library with respect to a binding function.
3. A method for identifying active or inactive
20 components of the peptide ladder library, i.e. identification of a protein signature for the protein or protein fragment under investigation with respect to the function being probed.

25 A combinatorial synthetic method for making a peptide ladder library is illustrated in Figure 1. The peptide ladder library is a one pot collection of

- 8 -

"n" peptides, each peptide being identical to the others in the library with respect to molarity and structure except for the substitution of a marker at position "n". The marker introduces a labile bond into the peptide backbone, e.g. a thioester bond, which can be selectively cleaved without cleaving other bonds within the peptide backbone. The marker also serves to introduce a ladder of stearic perturbations into the peptide backbone and/or to 5 introduce a ladder of peptide side chain 10 substitutions. The synthetic protocol employs a split synthesis method.

Conventional screening methods may be employed 15 on the peptide ladder library to separate active components from inactive components within the library. An exemplary screening protocol is 20 illustrated in Figure 2.

After the screening is complete, the isolated 25 components are analyzed as illustrated in Figure 3 to obtain a molecular signature for the protein. Briefly, the isolated components are cleaved at their marker and analyzed. Mass spectrometry is the preferred method of analysis. However, alternative analytical methods include nmr (with deuterium exchange), ir, and FACS. Comparison of the analysis, e.g., ms, of the isolate with the control, i.e., an aliquot of the entire library, provides a molecular 30 signature which identifies sites within the protein responsive or unresponsive to the screening method. For example, sites within the protein essential for

- 9 -

binding or folding may be identified. The protein signature of the Crk-N/C3G interaction is illustrated in Figure 3.

Successive iterations of the method of the invention can be employed to obtain a complete deconstructive analysis of a protein, even if the structure of the protein is unknown. The invention may be employed to characterize protein interactions and can facilitate the design of new therapeutics which are dependent upon such protein interaction.

One aspect of the invention is directed to a method for obtaining a molecular signature of a protein. The protein is of a type which has an amino acid sequence with length m , each amino acid position being represented by $(aa)_n$ where $1 \leq n \leq m$. The protein is also of a type which has a binding affinity with respect to a target molecule under binding conditions. The molecular signature then defined by a subsequence of the amino acid sequence of the protein. The subsequence is selected from amongst those positions $(aa)_n$ of the protein which, if individually replaced by a substitute amino acid, lead to a loss of binding affinity by the protein with respect to the target molecule.

The method employs a peptide ladder library. The peptide ladder library has m peptides. Each of the peptides is represented by $(peptide)_n$, where $1 \leq n \leq m$. Each peptide has the same amino acid sequence as the protein except that position $(aa)_n$ of

- 10 -

(peptide)_n is replaced by a substitute amino acid. Preferred substitute amino acids include alanine and glycine. If only one substitute amino acid is employed, then the peptide has a footprint the size of one amino acid. In alternative embodiments, the footprint may include two or three substitute amino acids. The substitute amino acid at position (aa)_n is linked to the amino acid at position (aa)_{n+1} by means of a labile bond. Preferred labile bonds are 5
thioester bonds and ester bonds. 10

The peptide ladder library is then contacted with the target molecule under binding conditions in order to form bound peptides and unbound peptides. The bound peptides are bound to the target molecule; 15
the unbound peptides are not. The unbound peptides are then separated from the bound peptides from said Step B in order to obtain separated unbound peptides. Each of the separated unbound peptide has the substitute amino acid only at position (aa)_n which 20
constitute the subsequence that define the molecular signature of the protein with respect to the target molecule. The labile bond of the separated unbound peptides are then cleaved in order to produce peptide cleavage products. Each peptide cleave product 25
corresponds to one of the positions (aa)_n from the subsequence which defines the molecular signature. The subsequence which defines the molecular signature of the protein is then constructed using the identity 30
of the peptide cleavage products to identify the subsequence of amino acid positions that are essential for binding to the target molecule.

- 11 -

Alternative substitute amino acids include the following: L-alanine, L-arginine, L-aspartic acid, L-asparagine, L-cysteine, L-cystine, L-glutamic acid, L-glutamine, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan, L-tyrosine, L-valine, D-alanine, D-arginine, D-aspartic acid, D-asparagine, D-cysteine, D-cystine, D-glutamic acid, D-glutamine, D-glycine, D-histidine, D-isoleucine, D-leucine, D-lysine, D-methionine, D-phenylalanine, D-proline, D-serine, D-threonine, D-tryptophan, D-tyrosine, D-valine, L- α -aminobutyric acid, D- α -aminobutyric acid, L- γ -aminobutyric acid, D- γ -aminobutyric acid, L- ϵ -aminocaproic acid, D- ϵ -aminocaproic acid, L-homophenylalanine, D-homophenylalanine, L-alloisoleucine, D-alloisoleucine, L- β -2-naphthylalanine, D- β -2-naphthylalanine, L-norvaline, D-norvaline, L-ornithine, D-ornithine, L-pyridyl alanine, D-pyridyl alanine, L-2-thienylalanine, D-2-thienylalanine L-methyltyrosine, D-methyltyrosine, L-citrulline, D-citrulline, L-homocitrulline, and D-homocitrulline.

In an alternative mode, the molecular signature of the protein is determined as described above except that the analysis is performed on the bound peptides are separated from the unbound peptides. Each of the separated bound peptides lacks any substitute amino acid at position (aa)_n from the subsequence which defines the molecular signature of the protein. The labile bonds of the separated bound peptides are then cleaved to form peptide cleavage

- 12 -

products. Each peptide cleave product corresponds to one of the positions $(aa)_n$ not included within the subsequence which defines the molecular signature. Accordingly, in this mode of the invention, after 5 detecting and identifying each of the peptide cleavage products, the subsequence which defines the molecular signature of the protein with respect to the target molecule is constructed by identifying amino acid positions $(aa)_n$ which do not correspond 10 to any of the peptide cleavage products.

Another aspect of the invention is directed a peptide ladder library corresponding to a protein. The protein is of a type which has a binding affinity with respect to a target molecule under binding 15 conditions. The protein is also of a type which has an amino acid sequence with length m where $1 \leq n \leq m$. Each amino acid position within the protein is represented by $(aa)_n$. The peptide ladder library then comprises m peptides, each peptide being 20 represented by $(peptide)_n$, where $1 \leq n \leq m$. Each peptide within the library has the same amino acid sequence as the protein except that position $(aa)_n$ of $(peptide)_n$ is replaced by a substitute amino acid. The substitute amino acid at position $(aa)_n$ is 25 linked to the amino acid at position $(aa)_{n+1}$ by means of a labile bond. If only one substitute amino acid is employed, then the peptide has a footprint the size of one amino acid. In alternative embodiments, the footprint may include two or three substitute 30 amino acids. Preferred labile bonds include thioesters and esters. Preferred substitute amino

- 13 -

acids are alanine and glycine.

Another aspect of the invention is directed to a method for constructing a peptide ladder library corresponding to a protein. The protein is of a type which has an amino acid sequence with length m . Each amino acid position of the protein may be represented by $(aa)_n$ where $1 \leq n \leq m$. The peptide library includes m peptides. Each peptide may be represented by $(peptide)_n$, where $1 \leq n \leq m$. Each peptide has the same amino acid sequence as the protein except that position $(aa)_n$ of $(peptide)_n$ is replaced by a substitute amino acid. The substitute amino acid at position $(aa)_n$ is linked to the amino acid at position $(aa)_{n+1}$ by means of a labile bond. A first reaction vessel may be provided which contains a first pool of nascent peptides having a length of $m-n$. The amino acid sequence of the nascent peptides runs between $n+1$ and m of the protein. The nascent peptides are attached to a matrix material. A second reaction vessel may be provided which contains a first pool of nascent ladder peptides having a length of $m-n$. The amino acid sequence runs between $n+1$ and m of the protein except that each $(nascent ladder peptide)_p$ has the substitute amino acid at position $(aa)_p$, where $n+1 \leq p \leq m$. The nascent ladder peptides are attached to a matrix material. An aliquot of matrix material is then transferred from the first reaction vessel to a third reaction vessel. Elongation reactions are then performed in each of the three reaction vessels. The first pool of nascent peptides in the first reaction vessel is

- 14 -

elongated by addition of the amino acid of position (aa)_n to form a second pool of nascent peptides having a length of $m-n+1$; the aliquot of nascent peptides in the third reaction vessel is then 5 elongated by addition of the substitute amino acid of position (aa)_n by means of labile bond to form a nascent ladder (peptide)_n having a length of $m-n+1$; and the first pool of nascent peptide ladders in the third reaction vessel is elongated by addition of the 10 amino acid of position (aa)_n to form a partial second pool of nascent peptide ladders having a length of $m-n+1$. After the elongation reactions are complete, the product of the third reaction vessel is transferred to the second reaction vessel to complete 15 the second pool of nascent peptide ladders having a length of $m-n+1$. The above process may then be repeated until $n=1$ and the second reaction vessel contains the sought after peptide ladder library.

20 Description of Figures:

Figure 1 represents the solid-phase peptide synthesis strategy used to scan a synthetic marker (dark oval) through sequential dipeptide units in a polypeptide sequence. Each member of the resulting peptide 25 family contains a single copy of the marker at a unique dipeptide site. By splitting and subsequently recombining the peptide-resin, all the members of the polypeptide family can be generated in a single synthesis. A modified solid phase peptide synthesis

- 15 -

methodology has been developed that makes it possible to prepare all members of the array of protein analogues concurrently in the course of a single synthesis. This simple procedure involves the use of two reaction vessels. At each stage of the synthesis a small aliquot of the peptide-resin is removed from the first vessel, and the analogue moiety attached to the growing peptide chain. The resin aliquot is then transferred to the second reaction vessel and the remainder of the amino acids in the sequence are coupled. Continual siphoning of resin aliquots from vessel #1 into vessel #2 (with analogue attachment in between), results in the generation of the complete protein array as a single product mixture. Use of this split-resin procedure ensures that each component of the array contains only a single copy of the analogue at a unique and defined position. The synthetic marker can be designed to probe the importance to structure and function of side-chain atoms, backbone atoms or both.

Figure 2 represents a one experiment cycle of iterative signature analysis. Multiple rounds of this cycle can be performed with the information from previous cycles being incorporated into each successive iteration.

Figure 3 illustrates preliminary iterative signature analysis data on residues 156-165 of the peptide Crk-N. a) Represents a comparison of the elution profiles obtained for the protein family using a C3G-peptide agarose column and a Leucine-Enkephalin agarose

- 16 -

column (control). Note that purified synthetic Crk-N binds the C33G-peptide in solution with a K_D of 2.3 μM (recombinant Crk-N = 1.9 μM). b) Represents HPLC profiles (all 25-45% acetonitrile over 30 minutes) obtained from the high salt wash (non specific) and ammonium acetate wash (specific) for the two columns. c) Represents the theoretical masses of peptide fragments produced upon ammonium acetate cleavage of a Crk-N family. Single letters refer to amino acid pairs substituted with the -Gly-SCH₂CH₂CO- marker. d) Comparison of the MALDI MS spectra of the ammonium acetate wash and the entire 9 component synthetic Crk-N family cleaved with ammonium acetate. In both cases the N-terminal peptide ladder is observed at a much higher intensity than the equimolar C-terminal peptide ladder.

Figure 4 illustrates the basic strategy for the synthesis of defined arrays of peptide analogues. The general approach is to have two main reaction vessels, one for unmodified peptide-resin, A, and the other for modified peptide-resin, B. Standard stepwise solid phase peptide synthesis of the parent amino acid sequence is performed in vessels A and B. Modifications to the sequence are made in a single auxiliary vessel, 1. At the beginning of each step in which introduction of an analogue structure is desired, a sample of peptide-resin is transferred from A to 1, where it is modified and then transferred from 1 to B after completion of that cycle of synthesis in both A and B.

- 17 -

Figure 5 illustrates the folding step for a family of peptides.

5 Figure 6 illustrates the screening method of the peptide ladder library with respect to binding function.

10 Figure 7 illustrates the readout step "unzipping the peptide" which reveals the latent chemistry.

15 Figure 8 illustrates an analysis of components using mass spectroscopy, nuclear magnetic resonance, HPLC, IR or FACS.

Figure 9 illustrates the process of the molecular signature analysis.

20 Figure 10 illustrates the design of a peptide ladder from the peptide sequence of a protein, protein fragment of bioactive peptide.

25 Figure 11 illustrates the scanning of a marker which introduces a labile bond into the peptide backbone. The labile bond can then be selectively cleaved without cleaving other bonds within the peptide backbone. The marker also serves to introduce a ladder of steric perturbations into the peptide backbone and/or introduce side chain substitutions.

Figure 12 illustrates a representative sample of the readout chemistry to introduce a labile bond into the peptide.

Figure 13 illustrates a process which comprises the

- 18 -

use of iterative steps as needed. The steps can be generally organized as 1) sequence info. 2) scan a perturbation 3) selection and 4) feedback.

5 Figure 14 illustrates an example of a peptide ladder library corresponding to a protein fragment of the SH-3 domain with the sequence -KGDILRIRDKP-.

10 Figure 15 illustrates: (A) the target composition of the nine member array of peptide analogues. The sequence PFKKGDILRIRDKPEE was derived from residues 152-167 of the murine cCrk SH3 domain and the C-terminal AcpRLKLKAR sequence was used to facilitate analysis by MALDI mass spectrometry; (B) Synthetic operations required for the synthesis of a peptide array consisting of nine overlapping dipeptide 15 analogues over a ten amino acid sequence. The synthesis was performed in a single day.

20 Figure 16 illustrates the analysis of a nine component array of peptide analogues. A) Analytical HPLC of crude full length product (gradient, 20-50% buffer B over 30 minutes). B) MALDI mass spectrum of crude full length product. Unlabelled peaks at lower mass are termination byproducts from the synthesis. 25 C) Analytical HPLC of hydroxylamine-cleaved HPLC product on the same gradient. D) MALDI readout of hydroxylamine-cleaved peptide array [Peaks with * are N-terminal-containing fragments; unlabeled peaks are C-terminal containing fragments].

Figure 17 illustrates the analytical HPLC of the chemical cleavage of model peptides containing labile

- 19 -

backbone bonds. **A)** Cleavage of a thioester-containing peptide with hydroxylamine. **B)** Cleavage of an ester-containing peptide by hydrolysis.

5 Figure 18 illustrates the schematic representation of the cleavage of the nine component array of peptide analogues. **A)** Full length array of nine peptide analogues. **B)** Cleaved array of peptide analogues. The mixture consists of eighteen peptides corresponding to nine N-terminal fragments and nine
10 C-terminal fragments.

Figure 19 illustrates the relation of the MALDI spectrum to the peptide C-terminal fragment array. The horizontal mass scale spectrum has been inverted to align it with the standard N-to-C terminal orientation of the peptide sequence. The peaks corresponding to the nine C-terminal peptide fragments are clearly resolved and can be assigned sequentially. In addition to the position of the peak in the mass spectrum, the mass difference
15 between adjacent peaks identifies the individual amino acids in the peptide sequence that has been subjected to analoguing. The starred peak corresponds to the callibrant.
20

Figure 20 illustrates the principle of protein
25 signature analysis. [1] Total chemical synthesis is used to generate an array of protein molecules derived from a single amino acid sequence. An analogue chemical structure (represented by the red-&-blue rectangles) is systematically incorporated

- 20 -

at defined positions in the polypeptide chain. [2] The array of protein analogues is subjected to functional selection, resulting in separation into two populations: active and inactive. [3] The 5 composition of each pool of analogues is then determined in a single step using a chemical readout system expressly built into the molecule for that purpose. This provides a signature relating the effects on function to substitution of the analogue 10 structure throughout the region of interest in the protein molecule.

Figure 21 illustrates an integrated strategy for the chemical synthesis, functional separation, and analysis of a self-encoded array of protein 15 analogues. The array of protein analogues is prepared by total chemical synthesis in a single procedure. Each analogue unit contains a selectively cleavable bond. Site-specific cleavage yields fragments that identify each protein component and 20 define the position of the analogue unit within the polypeptide chain. This decoding procedure is applied to the parent array of analogues, and to the active and inactive pools after separation based on function.

25 Figure 22 illustrates chemical Structures of Analogue Units A. Comparison of a native dipeptide unit (top) with the structures of the Gly-[COS]-Gly (middle) and Gly-[COS]- β Ala (bottom) analogue units used in the present study. B. Chemical cleavage of the thioester 30 bond within the analogue unit can be carried out

- 21 -

selectively under mild conditions by treatment with hydroxylamine at neutral pH. The thioester bond is stable to the conditions normally used to study proteins.

5 Figure 23 illustrates a readout of the composition of an array of analogues of the cCrk N-terminal SH3 domain. A. The array consists of nine sub populations, each containing a single Gly-S β Ala analogue unit. The dipeptide analogue was placed in
10 consecutive positions along the polypeptide chain, resulting in an overlapping pattern of substitution. B. Chemical cleavage of the thioester bond in the analogue unit results in an array of peptide fragments that characterizes the composition of the mixture of parent protein analogues. C. The array
15 of peptide fragments can be read out in one step by matrix assisted laser desorption time of flight mass spectrometry (MALDI-TOF). The resulting pattern of data is illustrated here for the C-terminal
20 containing family of fragments.

Figure 24 illustrates a combinatorial Readout of Protein Analogue Arrays. Building a latent chemical cleavage site into the analogue unit (rectangle) means that each protein in the array will contain this chemical marker at a unique position in the polypeptide sequence. Chemical cleavage specifically at the analogue unit gives rise to characteristic peptide fragments, each with a unique mass indicative of the position of the analogue unit within the sequence of the original protein analogue. Each
25
30

- 22 -

protein analogue in an array is thus self-encoded. Readout of these decoded peptide fragments can then be performed, in one operation, using MALDI mass spectrometry.

5 Figure 25 illustrates an application of protein signature analysis to a twenty residue region of the N-terminal SH3 domain of murine c-Crk. A. The highlighted amino acid sequence (residues 146-165) was substituted by the dipeptide analogue Gly-[COS]- β Ala, giving a 19-member array of synthetic analogue proteins. B. Signature obtained for the parent array, after cleavage with neutral hydroxylamine and analysis by MALDI mass spectrometry. Only the family of N-terminal fragments was observed in the spectrum.

10 The C-terminal fragments, although necessarily present, are not visible under the MALDI conditions used [Several terminated peptides, arising from impurities in the commercial amino acids used, are marked with an asterisk (*)]. C. Signature of the

15 active (binding) pool eluted from the C3G-derived synthetic peptide affinity column. Eight of the protein analogues displayed appreciable binding under these conditions, showing that dipeptide sequences N¹⁴⁶D¹⁴⁷; D¹⁴⁷E¹⁴⁸; E¹⁴⁸E¹⁴⁹; L¹⁵¹P¹⁵²; I¹⁶¹R¹⁶²; R¹⁶²D¹⁶³; D¹⁶³K¹⁶⁴; and,

20 K¹⁶⁴P¹⁶⁵ could be replaced by the dipeptide analogue without significant loss of activity. [Dipeptide sequences in parentheses, viz. '(ED)' & '(DL)', indicates the notable Gly-[COS]- β Ala-containing protein analogues not showing significant binding activity].

25

30

- 23 -

Figure 26 illustrates a readout of the composition of the parent nine component cCrk SH3 domain array, and the binding and non-binding pools.. The cCrk SH3 domain array of protein analogues was folded in assay buffer and added to a C3G peptide affinity column. Column fractions were treated with hydroxylamine and then analyzed by MALDI mass spectrometry. **A.** (control) Composition of the parent array of cCrk SH3 domain analogues. **B.** (Wash) Non-binding cCrk SH3 domain analogues eluted in the 0.5 M NaCl wash. **C.** (Elution) Specifically-bound cCrk SH3 domain analogues eluted with hydroxylamine. MALDI peaks are marked by the single letter code for the dipeptide that had been was substituted with Gly-S β Ala.

Figure 27 illustrates an iterative protein signature analysis applied to the N-terminal SH3 domain of murine c-Crk. **(Top)** The amino acid sequence of the 58 residue polypeptide chain is shown. Protein signature analysis was used to study how chemical variation of the centrally-located ten residue region (highlighted residues 156-165) affected C3G peptide binding. Two rounds of signature analysis were performed, using different dipeptide analogue units. **A. Round 1.** Signature of the active (binding) pool obtained from the nine-membered array of Gly-[COS]-Gly-containing protein analogues. In contrast to the previous experiment, analysis of this signature reveals that all nine protein analogues were present in the binding pool. **B. Round 2.** Signature of the active (binding) pool resulting from passing the parent array of Gly-[COS]- β Ala-containing protein analogues

- 24 -

over a C3G-derived synthetic peptide affinity column. The signature data shown represents an expansion of the larger signature shown in Figure 25C. Only four dipeptide sequences out of a total of nine (I¹⁶¹R¹⁶²; 5 R¹⁶²D¹⁶³; D¹⁶³K¹⁶⁴; and, K¹⁶⁴P¹⁶⁵) in this region, could be replaced by Gly-[COS]- β Ala without significant loss of binding activity.

Figure 28 illustrates a characterization of the purified synthetic murine cCrk 134-191, N-terminal 10 SH3 domain. A. Analytical HPLC of the total crude peptide products from HF cleavage. B. Analytical HPLC of the purified product on a gradient of 20%-50% B over 40 minutes. C. Electrospray mass spectra of the purified product. Inset spectrum is 15 reconstructed to a single charge state from the raw data below. Calculated mass for C₃₁H₄₇N₈O₉S₁ 6961.8 Da (average isotope distribution); Observed mass 6962 \pm 1 Da.

Figure 29 illustrates a characterization of the 20 parent array of synthetic analogues of the cCrk SH3 domain. A) Reverse phase HPLC analysis of the crude nine component protein array. B) MALDI mass spectrum of the same crude product mixture. The protein array contained predominantly full length polypeptide 25 products. The presence of lower molecular weight species in the MALDI spectrum result from termination reactions during chemical synthesis. A') Treatment of the array with hydroxylamine produces the cleaved peptide products. Reverse phase HPLC of the mixture 30 after chemical cleavage with NH₂OH showed partial

- 25 -

resolution of the 18 peptide fragments generated.
B') MALDI mass spectrometry of the same cleaved
mixture showed the characteristic patterns of
cleavage fragments. The peaks marked
5 (•) unambiguously identified the protein components
present in the original mixture. The order of these
peaks in the mass spectrum identifies the
corresponding analogue in the parent array and
defines the position of the analogue unit in the
10 polypeptide sequence.

Figure 30 illustrates affinity chromatography
performed on the nine component cCrk SH3 domain
protein analogue array as monitored by UV absorbance
at 280 nm. The 0.5 M NaCl wash eluted all non-
15 specific binding protein analogues as shown by
absence of significant elution with 1 M NaCl.
Protein analogues able to bind to the agarose-bound
C3G derived synthetic peptide were then eluted by
20 cleavage by hydroxylamine of the thioester bond in
each polypeptide chain. This procedure resulted in a
"functional selection" among the array of protein
analogues, giving a binding pool and a non-binding
pool. As a control, a column derivatized with a non-
specific peptide, derived from [Leu⁵] Enkephalin was
25 substituted for the C3G column. As shown, in the
hatched peaks, the entire SH3 protein array eluted
with the 0.5 M NaCl wash and no specific binding was
observed.

Figure 31 illustrates a region of the three-
30 dimensional structure of the c-Crk-C3G complex,

- 26 -

showing the three acidic residues within the RT loop of the SH3 domain interacting with Lys⁸ of the bound C3G peptide ligand. [Taken from a crystal structure]. These interactions are believed to make an important 5 contribution to binding, and to play a critical role in orienting the interaction of c-Crk with C3G.

Figure 32 illustrates results from two rounds of protein signature analysis of the sequence comprising residues 156-165 superimposed on the crystal 10 structure of the N-terminal c-Crk SH3 domain complexed to the proline rich C3G peptide. Indicated are those regions of the polypeptide chain observed to be either tolerant (green) or intolerant (red) of an extra backbone methylene group

Figure 33 represents a computer generated model of the SH3 domain. The green + red regions represent the perturbed sites of a peptide ladder fragment from a section of the SH-3 domain. The green molecular region represents that a perturbation in this region 15 has no effect on the activity with the ligand which is depicted in yellow. The red molecular region represents that a perturbation of this region 20 destroys activity with the ligand. The gray region represents the remaining unperturbed (unanalyzed) regions of the SH3 domain. The yellow molecular region represents a proline rich peptide ligand. 25

Detailed Description of the Invention

The invention is directed to a 3 step methodology, titled Protein Signature analysis, for

- 27 -

the identification of active or inactive components of a protein. The first step involves a combinatorial method for synthesizing a peptide ladder library corresponding to a protein, protein fragment, or other bioactive peptide. The second step comprises a method for screening the peptide ladder library with respect to a binding function. The third step comprises a method for identifying active or inactive components of the peptide.

5 Successive iterations of the method can be employed to obtain a complete deconstructive analysis of a protein, even if the structure of the protein is unknown. The invention may be employed to characterize protein interactions and can facilitate

10

15

the design of new therapeutics which are dependent upon such protein interaction.

The methodology combines the control of peptide composition provided by multiple synthesis of individual peptides with the synthetic convenience of a split and recombine synthetic strategy. By synthesizing an array of peptides which differ from a parent molecule by a limited number of defined modifications, the contribution of specific molecular features to peptide function can be probed in a systematic manner.

20

25

The methodology further comprises a novel encoding scheme which allows for the array of synthetic peptide analogues to be assayed free in solution. The composition of the peptide mixture can then be determined by a single readout operation.

30

- 28 -

The methodology was used to synthesize an array of peptide analogues in which a specific modification was systematically incorporated into unique positions in a peptide sequence (examples 1,2 and 3 infra).

5 The synthesis was carried out in such a way that the resulting mixture contained a defined family of modified peptides, with each peptide molecule containing only a single modification. The position of the analogue moiety within each member of the
10 array was self-encoded by incorporating a selectively cleavable bond into the analogue structure.

The synthetic polypeptide array was folded and analyzed for ligand binding on an affinity column as a single mixture, producing two separate binding and
15 non-binding pools of protein analogues.

Following selective cleavage of each polypeptide chain at the site of modification, the resulting mixture of peptide fragments (either the binding or non binding pool) was analyzed by MALDI mass spectrometry to generate patterns of data which defined the presence or absence of each peptide analogue in the ligand binding and non-binding pools. This mass spectrometric signature related the position of the chemical modification in the
20 polypeptide sequence to the ability to fold and/or bind to a specific ligand for the protein (examples 1,2 and 3, infra).

Example 1 (illustrates chemical synthesis and readout of self encoded arrays of peptide analogues - functional separation step is not included here)
30

- 29 -

A. Strategy for the Preparation of a Defined Array of Peptide Analogues as illustrated in Figure 4.

A peptide analogue array for use with the proposed self-encoding scheme has two important features. First, the components of the array must be present in approximately equimolar amounts. Second, to avoid ambiguities, the array should consist only of peptides containing a single chemical modification per peptide chain, at a defined number of positions in the sequence. A straightforward procedure for synthesizing an array of this type has been developed and is schematically represented in Figure 4. For simplicity, the procedure will be illustrated for a hypothetical array of peptides consisting of substitutions of a single amino acid analogue at each of ten consecutive positions in the amino acid sequence of the parent peptide.

Two manual solid phase peptide synthesis (SPPS) reaction vessels, A and B, and a small fritted funnel, 1, are used to manipulate the peptide-resin. The synthesis begins with ten units of peptide-resin in vessel A. After deprotection of the α -amino group, one unit of peptide-resin is removed from A and added to 1. The first amino acid is then coupled to the nine units of peptide-resin in A and the analogue moiety to the one unit peptide-resin sample in 1. After the coupling step, the analogue-modified peptide-resin from 1 is transferred to B.

- 30 -

To initiate the next cycle of synthesis, the peptide-resin in vessels **A** and **B** are deprotected. Then another unit of peptide-resin is removed from **A** and transferred to the now empty **1**. The next amino 5 acid in the sequence of the parent peptide is added in activated form to both **A** and **B**, while the analogue moiety is reacted with the new peptide-resin sample in **1**. After completion of this cycle, the modified peptide-resin in **1** is added to **B**. The synthesis 10 continues in this manner for the requisite ten cycles.

Throughout the synthesis, vessel **A** contains only unmodified peptide-resin. Vessel **B** contains all single-site modified peptide-resins and vessel **1** 15 contains the current sample of peptide-resin which is being modified. All chemical steps carried out in vessels **A** and **B** are identical, adding the amino acids of the unmodified sequence. At the end of 10 cycles, all the resin in vessel **A** has been transferred into 20 vessel **B** which now contains the desired array of peptide analogues in resin-bound form.

B. Synthesis of a Defined Peptide Array.

A peptide array consisting of a ten amino acid sequence, GDILRIRDKP was chosen as a target to 25 demonstrate the approach, following the methodology as described above. The target array is shown in Figure 15A, and consists of overlapping dipeptide analogues in the region of interest. In order to facilitate characterization by mass spectrometry, the

- 31 -

array was synthesized on resin bearing the sequence EEAc_{PLK}KLKAR, where Ac_p is ϵ -aminocaproic acid (Zhao et al. Proc. Nat. Acad. Sci. 1996, 93, 4020-4024). The dipeptide analogue moiety corresponding to -NH-
5 CH₂-CO-S-CH₂-CH₂-CO- (Gly-S β Ala) was introduced as Boc-Gly-S β Ala. Since the analogue moiety was incorporated as a dipeptide, a modification was made to the synthetic procedure outlined above and shown in Figure 4. In order to keep the synthetic
10 operations being performed on the peptides in vessels A and B in register, the sample being derivatized in 1 was held out for two cycles before transfer to vessel B. To accommodate this modification, a second auxiliary funnel 1' was added.

15 In practice, the peptide-resin sample from vessel A was added to a funnel in position 1, where the dipeptide analogue coupling was initiated. After one cycle, the funnel was moved to position 1', where the dipeptide analogue coupling continued during a
20 second cycle of chain elongation in vessels A and B. The analogue-containing sample of peptide-resin was then washed with DMF (dimethylformamide) and transferred to vessel B. The synthetic steps for this synthesis are outlined in Figure 15B. After
25 substituting dipeptide analogues for nine consecutive dipeptide sequences spanning a region of 10 amino acids, four additional amino acids, PFKK, were coupled to the array of peptide-resins in vessel B to complete the target sequence.

30 C. Characterization of the Peptide Array.

- 32 -

The peptide array described above contains a mixture of nine peptides, all 24 residues in length, each differing only in the position of a Gly-S β Ala dipeptide substitution. As expected, the analytical HPLC of this array is quite complex, with many overlapping peaks (Figure 16A) The MALDI mass spectrum is also poorly resolved since the peptides in the array have a high redundancy in their molecular weights (Figure 16B). Thus the sequence -
5 LRIRD- contains the dipeptides LR, RI, IR, each of which have a molecular weight of 269 Da and RD which has a molecular weight of 271 Da. When substituted with Gly-S β Ala (145 Da) in the peptide arrays, each of these substitutions would result in a peptide
10 analogue with a molecular weight of 125 \pm 1 Da below that of the unmodified sequence. (PFKK-GDILRIRDKP-EEAcpRLKLKAR·amide, M.W. 2920 Da). The resulting MALDI mass spectrum of this peptide array would be expected to have a large peak around 2795 Da,
15 representing the sum of four different peptide components, (see Figure 16B).
20

D. 'Self-Encoded' Peptide Arrays.

The poor HPLC separation and redundancy in molecular weight creates a challenge for
25 identification of components present in the array of peptide analogues. The distinguishing feature of the components in this peptide array is the unique position of the modification in the sequence. One approach to the unambiguous identification of the
30 peptide components is to incorporate a selectively

- 33 -

cleavable bond in the analogue unit. Cleavage of this bond in an analogue peptide would result in two peptide fragments whose lengths, measured as mass, would define the position of the analogue unit in the 5 peptide from which they derived. Such a chemical cleavage site would have to be stable to normal handling (folding and assay) conditions, while permitting selective cleavage on demand. We have investigated the incorporation, stability and 10 selective cleavage properties of two potential readout chemistries, based on ester and thioester bonds.

15 1. Synthesis and Characterization of a Peptide Containing a Cleavable Thioester Backbone as illustrated in Figure 17A.

The peptide LYRA(Gly-S β Ala)-YGGFL·amide, was synthesized by stepwise SPPS using *in situ* neutralization coupling protocols. The thioester-containing dipeptide analogue, Boc-Gly-S β Ala was activated as an HOEt ester and then coupled to pre-neutralized NH₂-YGGFL-(4-Me)benzhydrylamine)-resin. Following deprotection and cleavage from the peptide-resin, the stability of the model thioester-containing peptide was determined.

25 Thioester bonds within peptide sequences have been found to be stable at neutral pH (Schnölzer et al. *Science* 1992, 256, 221-225; Baca et al. *J. Am. Chem. Soc.* 1995, 117, 1881-1887; Canne et al. *J. Am.*

Chem. Soc. 1996, 118, 5891-5896). To test for stability to base hydrolysis, the peptide was dissolved at pH 9.0 in 200 μ L of 100 mM Tris, 1 M Gn·HCl, vortexed vigorously for 10 seconds and left at 23°C for 30 minutes. Surprisingly, no hydrolysis was observed under these conditions. Addition of 20 μ L of 1 M NaOH to (pH~13) gave complete hydrolysis after just 10 minutes as monitored by HPLC and electrospray mass spectrometry. In contrast to their stability to hydrolysis, thioesters have been shown to be very labile to hydroxylamine at neutral pH levels (Bruice et al. *J. Am. Chem. Soc. 1964, 86, 4886-4897*). As shown in Figure 17A, the thioester peptide was completely cleaved into LYRAG-NH₂OH and HSCH₂CH₂CO-YGGFL·amide when dissolved in 1 M NH₂OH, 200 mM NH₄HCO₃, pH 6.0 for 30 minutes. Thioesters can be completely cleaved at concentrations of NH₂OH as low as 10 mM in < 30 minutes.

2. Synthesis and Characterization of a Peptide Containing a Cleavable Ester Backbone as illustrated in Figure 17B.

The peptide, YKLFAla-[coo]-YGGFL·amide was prepared by stepwise SPPS, using *in situ* neutralization coupling protocols and Boc chemistry. The ester bond in the peptide was formed by coupling Boc-Ala to an α -hydroxy acid using 4-dimethylaminopyridine as a catalyst. Following deprotection and cleavage from the peptide-resin, the model ester-containing peptide was analyzed for

- 35 -

stability. The ester bond was quite resistant to hydrolysis, taking six hours to cleave at pH 10, Figure 17B. In addition, the ester peptide was stable to treatment with 1 M NH₂OH, 200 mM NH₄HCO₃, pH 6.0 for up to 12 hours. The high stability of the ester to hydroxylamine should allow for the use of a backbone thioester as a chemical readout in peptides containing ester bonds. More recent studies have shown that the ester can be readily cleaved by NH₂NH₂ at neutral pH. A Gly-[coo]-Gly containing peptide was cleaved in under one hour by dissolving in 150 mM hydrazine, 100 mM Sodium Phosphate, 6 M guanidine·HCl, pH 7.0 (Carrasco, M. unpublished).

E. Readout of the nine component peptide
15 analogue array of the parent sequence PFKK-
GDILRIRDKP-EEAcpRLKLRKAR·amide.

Synthesis of this array has been described above. Each member of this peptide array contains a Gly-S β Ala replacement at one of the nine possible 20 dipeptide positions within the ten amino acid sequence, [GDILRIRDKP] as shown in Figure 15A. In order to facilitate unambiguous identification of the components in the array, the thioester readout chemistry has been incorporated into the dipeptide 25 analogue. The thioester bond in the Gly-S β Ala dipeptide analogue introduces a unique cleavage site into each member of the peptide array. Chemical cleavage of the peptide analogue array was expected to produce 18 peptide fragments as shown in Figure 30 18.

- 36 -

The peptide analogue array was cleaved by treatment with 1 M NH₂OH, 200 mM NH₄HCO₃, pH 6.0 for 20 minutes. The resulting peptide fragments were then analyzed by HPLC and by MALDI mass spectrometry.

5 As with the uncleaved peptide array, Figure 16A, the components of the cleaved peptide array, Figure 16C, still give rise to a complicated and essentially uninformative HPLC chromatogram. By contrast, the MALDI spectrum of the unfractionated cleaved peptide

10 array, provides a very straightforward characterization of the peptide array, Figure 16D. As shown in Figure 19, the masses of the nine C-terminal fragments from the cleaved array are easily located in the mass spectrum (Signals corresponding

15 to 5 of the 9 N-terminal peptides were also observed. The masses that were not resolved were obscured by matrix ions in the region below 1000 Da in the spectrum of the peptide mixture), and served to unambiguously identify the position of the

20 analogue unit in each component of the original peptide array before cleavage.

The MALDI mass spectrometric readout characterized the peptide array in several ways. Cleavage of the analogue unit, at different positions

25 throughout the array, produces two families of related peptides. These two families have either the N- or C-terminus of the parent peptide in common. In this experiment, the identification of each member of the C-terminal family of peptides by MALDI can be

30 directly related to the presence of each full length peptide analogue in the parent array. In addition to

- 37 -

the C-terminal family, five members of the N-terminal family of peptides were also observed. The starred peaks on mass spectrum in Figure 16D correspond to these N-terminal peptides. In practice, the 5 identification of either the N-terminal or C-terminal peptide fragments would serve to unambiguously characterize the peptide analogue array.

Another characterization of the peptide array can be obtained by looking at the difference in 10 masses between the peaks on the mass spectra. As shown in Figure 19, the nine C-terminal peptide fragments are all of different lengths, and the mass differences between neighboring fragments correspond to the masses of individual amino acid residues. By 15 correlation of these mass differences to the mass of individual amino acids, the sequence through which the analogue unit was substituted can be confirmed. This type of analysis has been previously used to sequence native peptides by 'protein ladder 20 sequencing' (Chait et al. *Science*. 1993, 262, 89-92).

F. Discussion

An embodiment of the present invention is the synthesis of a defined array of peptide analogues as a single mixture. This peptide array can be assayed 25 as a pool, after which individual peptide components can be identified through a novel encoding scheme. Using this approach, the comparative properties of the individual components can be determined for a given assay/function. The adaptation of this

analogue array methodology to the study of structure-function relationships in proteins is illustrated in Example 2 (infra).

The peptide array can be synthesized using a modified 'split resin' procedure which, unlike previous procedures, results in a defined array of components with only the desired complexity. A key aspect of this approach is the ability to prepare a defined subset of a fully combinatorial synthesis.

By synthesizing such a subset, arrays of manageable size can be prepared that contain defined modifications at a large number of positions in the peptide sequence. For example, using this methodology, five different amino acid analogues could be substituted at each position of a ten amino acid sequence so that there is only one modification in each peptide molecule. The resulting array would result in a mixture of 50 peptides (5 analogue structures x 10 positions). Using the standard split synthesis approach (Furka et al. *Int. J. Pept. Prot. Res.* **1991**, 37, 487-494; Lam et al. *Nature* **1991**, 354, 82-84), the same 50 peptides would be only a small fraction of a library of $\sim 10^7$ (5^{10}) peptides.

Once diversity has been generated, and a selection performed, a method is needed to identify individual components. Such approaches to decoding peptide mixtures have presented a substantial challenge. Most encoding strategies involve a molecular tag which can be read by sensitive analytical techniques (Needels et al. *Proc. Natl.*

- 39 -

Acad. Sci. U.S.A. 1993, 90, 10700-10704; Kerr et al.
J. Am. Chem. Soc. 1993, 115, 2529-2531; Nikolaiv et
al. *Pept. Res.* 1993, 6, 161-170; Ohlmeyer et al.
Proc. Natl. Acad. Sci. U.S.A. 1993, 90, 10922-10926)
5 or by amplification (Brenner et al. *Proc. Natl. Acad.
Sci. USA.* 1992, 89, 5381-5383; Nielsen et al. *J. Am.
Chem. Soc.* 1993, 115, 9812-9813). Many of these
techniques, however, rely on the assay of molecules
still attached to a solid support and the isolation
10 and analysis of individual beads. To avoid the
necessity for a solid support, the encoding must be
associated with the peptide analogue at the molecular
level (Brenner et al. *Proc. Natl. Acad. Sci. USA.*
1992, 89, 5381-5383; Nielsen et al. *J. Am. Chem.
Soc.* 1993, 115, 9812-9813).

Incorporation of a chemically cleavable bond at
specific sites within the peptide analogues provides
an example of an alternative simple and practical
encoding scheme at the molecular level. After a
20 physical selection for a desired function, the
components of the peptide array can be decoded
through chemical cleavage and one step mass
spectrometric readout. Detection of the resulting
peptide fragments unambiguously defines the presence
25 or absence of a given analogue molecule in the
selected population. As demonstrated in the analysis
of the nine member peptide array, matrix assisted
laser-desorption ionization (MALDI) mass spectrometry
is well suited for the decoding of linear arrays of
30 peptides (Chait et al. *Science.* 1993, 262, 89-92).

- 40 -

This use of mass spectrometry to decode mixtures of peptide analogues is analogous to the use of gel electrophoresis to separate nucleotides by length during DNA sequencing and analysis (A similar 5 cleavage and separation by size readout was used for nucleic acids in: Pan et al.; *Science*. 1991, 254, 1361-1364; Hayashibara et al. *J. Am. Chem. Soc.* 1991, 113, 5104-5106). The high resolution and 10 sensitivity (<1 pmol/component; Chait et al. *Science* 1992, 257, 1885-1894) of MALDI mass spectrometry allows the characterization of even small quantities 15 of the entire peptide array.

In this example, we have demonstrated the 20 feasibility of this approach through the synthesis of a peptide array which was then characterized by MALDI mass spectrometry following chemical cleavage. A nine component peptide array of the parent sequence, PFKK-GDILRIRDKP-EEAc_nRLKLKAR·amide, was synthesized in which a single dipeptide analogue, Gly-S β Ala, was 25 introduced into consecutive positions through the sequence -GDILRIRDKP- (Figure 15). This array was self-encoded with a chemically cleavable thioester bond which was incorporated into the analogue unit. The peptide components were then identified by 30 cleaving the thioester bond in each peptide with hydroxylamine, followed by MALDI mass spectrometry. The resulting series of peaks on the mass spectrum unambiguously identified the presence of all nine peptide analogues in the peptide array.

30 The combination of the synthesis of an array of

- 41 -

peptides corresponding to a defined subset of a fully combinatorial mixture with a self-encoding strategy results in an information-rich approach to the elucidation of the structure-activity relationship of 5 peptides, polypeptides and proteins. The power of this approach is illustrated when the peptide array is subjected to a functional selection (example 2). Since all members of the peptide array can be observed in a single readout step, this approach can 10 generate information on both the positively and negatively selected members of the array. It is anticipated that the ability to synthesize multiple peptide analogues in a single procedure, followed by functional characterization of the entire peptide 15 array will give greater insight into the molecular basis of peptide function.

This procedure allows for controlled diversity to be generated at multiple positions in a peptide chain. In addition, a novel self-encoded approach to 20 identifying peptide analogues has been developed which involves the incorporation of a cleavable bond which is associated with a particular modification. This chemical readout system allows the entire peptide array to be analyzed simultaneously by 25 sensitive analytical techniques such as MALDI mass spectrometry.

By reading out an entire pool of peptide analogues in a single step, a profile of structure-function relationship across all members of the array 30 could be generated. Interpretation of such profiles

- 42 -

will provide information of the molecular basis for peptide function. Such peptide arrays could be used to elucidate the molecular basis of important functional properties by systematically removing structural elements of a peptide. In particular, hydrogen bond donors and acceptors in the backbone can be deleted using a wide variety of backbone analogues. Additionally, new functional characteristics can be introduced through the systematic introduction of new sidechain and backbone groups. The use of defined arrays of synthetic peptide analogues coupled with a single step readout provides new insights into the chemical basis of peptide activity.

15

Example 2 (illustrates chemical synthesis, functional separation, and readout of self-encoded arrays of an SH3 peptide consisting of the ten amino acid sequence, GDILRIRDKP)

20 This example illustrates an integrated approach to the preparation of a defined array of protein analogues in a single synthesis, their functional separation into active and inactive pools, and a simple one step readout of the composition of the self-encoded mixtures. The strategy is outlined in Figure 21. The chemical synthesis and encoding strategies are an extension to proteins of the approach described in example 1 (supra). Instead of synthesizing a single modified protein, an array of protein analogues is prepared in a single total

30

- 43 -

chemical synthesis. The proteins in the array differ from each other only by the position in which a defined covalent modification is located within the polypeptide sequence. The composition of the 5 analogue array can be decoded by means of a latent readout chemistry introduced in conjunction with each analogue unit. This latent readout chemistry allows proteins to be specifically cleaved, yielding a pattern of peptide fragments which unambiguously 10 identifies the individual components of the mixture and defines the position of the analogue unit in each compound (A similar readout system has been developed for use in DNA systems: Hayashibara et al. *J. Am. Chem. Soc.* 1991, 113, 5103-5106). By comparing the 15 readout patterns obtained before and after a functional separation, a profile is obtained relating the effects of the analogue structure to its position in the polypeptide chain. Accumulation and interpretation of such qualitative profiles of 20 protein structure-function relationships 'protein signatures' provides insights into the chemical basis of protein function.

A family of analogue proteins was characterized 25 using the invention with the N-terminal SH3 domain of the adapter protein cCrk (Murine cCrk, N-terminal SH3 domain (residues 134-191); Genbank accession s72408). SH3 domains are small monomeric modules which are present in many proteins involved in signal 30 transduction. It is now well established that SH3 domains mediate protein-protein interactions within intracellular signaling networks through the

- 44 -

recognition of short proline-rich sequences from other adapter proteins (Ren et al. *Science* 1993, 259, 1157-1161). Individual SH3 domains have been shown to fold *in vitro* to a defined tertiary structure and to bind proline-rich peptides with low μ M affinity. In addition several of these domains have been structurally characterized by both NMR and X-ray crystallography (Musacchio et al. *Nature* 1992, 359, 851-855; Yu et al. *Science* 1992, 258, 1665-1668).

Using the methods described in example 1 (supra), a family of nine analogues was prepared to investigate the effect of an extensive covalent modification of the 58 amino acid residue SH3 polypeptide chain on its ability to fold and bind its specific peptide ligand. Each member of the synthetic protein array contained a single dipeptide analogue unit, $-\text{NH}-\text{CH}_2\text{CO}-\text{SCH}_2\text{CH}_2\text{CO}-$ (Gly-S β Ala), replacing pairs of adjacent amino acids at unique positions in the native sequence (see Figure 23). The mixture of analogue polypeptide chains was folded and assayed for binding to a specific ligand, a short proline-rich synthetic peptide derived from the sequence of the guanine nucleotide exchange factor C3G (Knudsen et al. *J. Biol. Chem.* 1994, 269, 32781-32787). After the affinity selection, the binding and non-binding pools of protein analogues were cleaved selectively at the thioester bond contained in each analogue unit, and the composition of each pool was read out using MALDI mass spectrometry. In this manner, a pattern of signals was obtained, which related the position of the chemical modification

- 45 -

within the SH3 polypeptide sequence to its effects on folding and/or ligand binding.

A. Chemical Synthesis of the 58 residue SH3 polypeptide

5 The polypeptide chain of the murine cCrk N-terminal SH3 domain, corresponding to residues 134-191 of the full cCrk signaling protein (Murine cCrk, N-terminal SH3 domain (residues 134-191); Genbank accession s72408), was assembled by highly optimized, 10 stepwise solid phase peptide synthesis using machine-assisted *in situ* neutralization protocols for tert-butoxycarbonyl (Boc) chemistry (Schnölzer et al. *Int. J. Pept. Protein Res.* 1992, 40, 180-193). After deprotection and cleavage from the resin support, the 15 crude polypeptide was purified by semipreparative reversed-phase HPLC and lyophilized using procedures as described by Clark-Lewis et al. *The Use of HPLC in Receptor Biochemistry*; Venter, J.C. and Harrison, L.C. Eds.; Alan R. Liss, Inc: New York, 1989; Chapter 3. The purified product was characterized by 20 analytical HPLC and by electrospray mass spectrometry (Schnölzer et al. *Anal Biochem.* 1992. 204. 335-343). The results are shown in Figure 28.

B. Functional characterization of the synthetic SH3 domain

25 The N-terminal cCrk SH3 domain was formed by folding under these conditions: 0.2 mg of the purified 58 residue polypeptide in 600 μ L of 20 mM

- 46 -

HEPES 50 mM NaCl pH 7.3 at room temperature for 15 minutes. The folded protein was structurally characterized by NMR and crystallization (The folded protein was characterized by two dimensional NOESY 1H NMR spectroscopy. In addition, the synthetic protein solution used for NMR analysis spontaneously formed crystals upon storage at 4°C. These crystals diffracted to >2.3 Å allowing for the solution by molecular replacement. Refinement of the structure is in progress). The resulting synthetic protein domain was then assayed for its affinity for two different proline-rich peptides. Binding of this SH3 domain to its cognate peptide ligand buries a tryptophan sidechain of the protein found in the binding pocket. This change in solvent exposure leads to an increase in the fluorescence intensity of the tryptophan sidechain which can be monitored as a function of increasing ligand concentration (Feng et al. *Science* 1994.266, 1241-1245). The K_a's for the peptides C3G (PPALPPKKR·amide) and a peptide designed for attachment to an affinity column, Acetyl-CWAcp-C3G, were found to be 1.8 μM and 2.4 μM respectively (Due to the polyproline nature of the ligands, fluorescence measurements were taken after 12 hours 25 incubation of the protein with the peptide ligand to allow equilibration of the multiple cis and trans isomers). The C3G affinity of 1.8 μM for the C3G derived peptide is comparable to the affinity reported for a recombinantly derived SH3 domain (1.9 μM) (Wu et al. *Structure* 1995, 3, 215-226).

C. Affinity chromatography of the synthetic

SH3 domain

An affinity column carrying a synthetic C3G derived peptide was prepared by reaction of the cysteine side chain of Acetyl-CWAcP-PPALPPKKR·amide with an iodoacetylated agarose matrix. The linker Acetyl-CWAcP- was designed to combine a reactive sulfhydryl on the Cys sidechain with a spectroscopic tag in the Trp residue, and to introduce a flexible spacer region between the support and the peptide ligand in the form of ϵ -aminocaproic acid. The amount of peptide on the column, was determined as 5 μ mol/mL of the swollen agarose matrix from the absorbance at 280 nm of the peptide solution before and after reaction with the column. As a control for non-specific binding effects, a second column with comparable peptide loading was made by the same procedure using a non C3G sequence, Ac-CWAcP γ GGFL·amide.

Specific binding of the native sequence synthetic cCrk N-terminal SH3 domain was demonstrated to the C3G-peptide affinity support by loading crude synthetic cCrk(134-191) mixed with BSA. The proteins were allowed to equilibrate for 6 hours after which a series of increasing washes up to 1 M NaCl, and one wash with 1 M NH₂OH was performed. The column effluent was monitored by absorbance at 280 nm and selected fractions subjected to mass spectrometry. The BSA was completely washed off the column after the 300 mM NaCl wash. The cCrk N-terminal SH3 domain, however, remained bound to the column

- 48 -

throughout the NaCl and NH₂OH washes. The cCrk domain was only eluted by a 6M GnHCl wash that disrupted the specific interactions between the protein and the synthetic C3G-derived peptide. It is interesting to 5 note that the 58 amino acid cCrk SH3 domain was stable to 1 M NH₂OH treatment despite the presence of the sequence Asn-Gly-Asn (144-146) which can be prone to NH₂OH cleavage (Clarke et al. *Stability of Protein Pharmaceuticals, Part A:*, Ahern, T. J.; Manning, M.C. 10 Ed. Plenum Press, New York, 1992).

D. Chemical synthesis of an array of analogues of the 58 residue polypeptide chain

The target array consisted of nine polypeptides, each containing a single -NHCH₂CO-SCH₂CH₂CO- (Gly-S β Ala) substitution at one of the nine possible 15 dipeptide units within the ten amino acid sequence defined by residues 156-165 within the cCrk sequence 134-191 (Figure 23). The Gly-S- β Ala analogue unit 20 was designed to remove two consecutive amino acid sidechains and to insert an extra methylene into the polypeptide backbone of the SH3 protein. Inclusion 25 of the thioester bond added a latent chemical cleavage site for the readout of the composition of the mixture of protein analogues and simultaneously deleted a backbone hydrogen bond donor. Backbone flexibility was also increased due to the loss of the planarity associated with the peptide bond. The 30 analogue substitutions covered overlapping dipeptide sequences through the region -GDRILRIDKP-, corresponding to residues 156-165 of the cCrk

- 49 -

sequence as shown in Figure 23A.

The polypeptide analogue array was synthesized using a combination of manual and machine-assisted protocols. The sequences corresponding to cCrk(166-191) and cCrk(134-155) were synthesized using a machine-assisted protocol for Boc-solid phase chemistry (Schnölzer et al. *Int. J. Pept. Protein Res.* 1992, 40, 180-193). Following synthesis of cCrk(166-191), the peptide-resins were removed from the peptide synthesizer and placed in a manual-synthesis reaction vessel. Synthesis of cCrk(156-165) with concomitant introduction of the analogue unit at each position was performed as previously described for a model peptide system (A similar readout system has been developed for use in DNA systems: Hayashibara et al. *J. Am. Chem. Soc.* 1991, 113, 5103-5106) using a modified split-resin procedure. Before addition of each residue (157-165), a sample of peptide-resin was removed from reaction vessel A, modified with a dipeptide analogue for two synthetic cycles and then transferred to a second reaction vessel B. Identical synthetic operations building up the native amino acid sequence were carried out in reaction vessels A and B. In this manner, an array of nine resin-bound polypeptides was created as a single mixture, containing consecutive overlapping Gly-S β Ala substitutions. Chain elongation of the nine component mixture was continued through the sequence cCrk(134-155) using machine assisted synthetic cycles. The mixture of full-length analogue-

- 50 -

containing polypeptide-resins was subjected to HF cleavage and simultaneous sidechain deprotection to give a crude lyophilized mixture of nine analogues of the 58 amino acid polypeptide chain of the N-terminal cCrk SH3 domain.

The members of the polypeptide array were folded by dissolving 0.5 mg of crude peptide product in 200 μ L of 20 mM HEPES 50 mM NaCl pH 7.3 at room temperature. After 1 hour, the protein mixture was analyzed by HPLC and MALDI mass spectrometry. By HPLC (Figure 29A), the nine protein components were partially resolved against a background of synthetic byproducts. MALDI mass spectrometry (Figure 29B) showed an unresolved mixture of full-length polypeptide chains centered around 6950 Da; minor amounts of terminated components formed as byproducts in the chain assembly were also present. The presence of full length protein analogues indicates that the thioester-containing polypeptide chains are stable under these assay conditions. However, neither HPLC nor MALDI-MS was able to define the composition of the array of protein analogues.

E. Readout of the Composition of the Parent Array of Protein Analogues

In order to characterize the protein array, the polypeptide chains were specifically cleaved with hydroxylamine through nucleophilic attack at the thioester bond. The resulting peptide fragments were analyzed by both HPLC and MALDI-TOF, Figures 29A' and 29B'. The treatment with hydroxylamine specifically

- 51 -

cleaved each analogue-containing polypeptide chain at the site of modification, resulting in a mixture of peptide fragments. As shown in Figure 23, the cleavage of the protein array produces a mixture of peptide fragments, each with a different number of amino acids in the peptide chain. As shown in Figure 29A', reverse-phase HPLC analysis of the cleaved mixture yields complicated sets of partly unresolved peaks. In addition, the order of the peaks bears no direct relationship to the position of the analogue unit in the parent polypeptide chain. Analysis of the cleaved mixture by MALDI mass spectrometry, on the other hand, produces a series of well resolved peaks, the relative positions and masses of which are directly related to the position in which the analogue unit was placed in each full-length polypeptide chain. The peptide ladder shown in Figure 29B' corresponded to all nine of the expected N-terminal peptide fragments resulting from cleavage of the nine protein analogues, and unambiguously characterized the protein array.

Since the cleavable bond was placed within the 58 residue SH3 polypeptide chain, the cleavage of each analogue must produce both N- and C-terminal peptide fragments. In the MALDI spectrum, however, only the N-terminal peptide fragments were observed with high intensity. The analysis of peptide mixtures by MALDI mass spectrometry can vary depending on the choice of matrix and solvent composition. This phenomena did not compromise our experimental results since the MALDI readout system

- 52 -

relied on the comparison of mass spectra before and after selection. In addition, detection of only one of the two peptide fragments from each protein was required for identification of the parent protein analogue. In fact, interpretation can be simplified when only fragments corresponding to one end of the polypeptide chain are observed.

F. Affinity chromatography of the array of cCrk SH3 protein analogues

The lyophilized crude mixture of 58 residue polypeptide chain analogues was folded by dissolving 1.5 mg in 600 μ L of 20 mM HEPES, 50 mM NaCl, pH 7.3. The dissolved protein array was then applied to the affinity column and left to bind for 6 hours. The column was then washed with 0.5 M NaCl buffer to remove nonspecific binding proteins from the column. Specifically bound protein analogues were then eluted with 1 M hydroxylamine buffer. Amounts of eluted peptide were monitored by UV absorbance at 280 nm. This procedure was used for both the C3G peptide column and for the control column loaded with Ac-C-AcpYGGFL-amide. The hydroxylamine cleavage of the thioester bond results in the elution of peptide fragments corresponding to the proteins which were able to bind to the affinity column under these conditions. The results obtained are shown in Figure 30. Specific binding was found only for the C3G peptide column.

G. Readout of the composition of the arrays of protein analogues

- 53 -

The composition of the parent array of SH3 protein analogues, as well as the binding and non-binding fractions obtained from affinity chromatography was determined by chemical cleavage and MALDI MS. First, the parent array of folded protein analogues and the 0.5 M NaCl wash from the affinity chromatography of the array were cleaved separately with hydroxylamine. After a desalting step, MALDI mass spectra were obtained for the peptide fragments generated from the (parent) protein array, the salt wash (non-binding) and for the peptide fragments generated by the hydroxylamine elution of the specifically bound analogues (binding). The results are shown in Figure 26. Nine components were present in the array of protein analogues that was added to the C3G column as one mixture (Figure 26A). Of these nine components, five did not bind significantly to the affinity column and were present only in the 0.5 M NaCl wash fractions (Figure 26B). Three protein analogues, however, were able to bind to the C3G peptide on the column and were eluted only after cleavage with hydroxylamine (Figure 26C). One of the protein analogues can be identified in both the wash and elution spectra, indicating intermediate folding and/or binding properties for this analogue.

The affinity assay used in this experiment did not allow binding effects to be distinguished from folding effects since lack of binding to the affinity column could arise either from failure to fold or from correctly folded material failing to bind. The

- 54 -

observed pattern of functional and non-functional analogues covering the sequence 156-165 of cCrk indicated that in four of the nine positions even extensive modifications to the chemical structure of 5 the polypeptide chain were not sufficient to prevent folding and specific ligand binding.

H. Discussion

The present invention is the first application 10 of combinatorial synthetic chemistry techniques to a protein target. Straightforward synthetic access to protein arrays containing a dipeptide analogue Gly-S β Ala has been demonstrated in the context of residues 156-165 of the cCrk N-terminal SH3 domain (residues 134-191) of cCrk. A latent readout 15 functionality has been shown to be stable to conditions of protein folding and ligand binding, yet is cleavable by brief treatment with 1 M hydroxylamine. A selection for binding activity based on the use of a C3G-derived synthetic peptide 20 for affinity chromatography has been developed and used to analyze the functional properties of the array of protein analogues. Finally, chemical cleavage of synthetically introduced latent cleavage sites followed by MALDI-TOF mass spectroscopy has 25 been used to read out the composition of 'self-encoded' pools of protein analogues.

The analogue unit used in this study caused the simultaneous modification of several aspects of the covalent structure of the polypeptide chain. The 30 Gly-S β Ala dipeptide consisted of four alterations

- 55 -

from the native dipeptide; two sidechain deletions, an extra methylene in the backbone and a thioester substituted for the amide bond. By redesigning the analogue unit in an iterative manner, further insight
5 into the individual components of the modification can be elucidated. Although the thioester is needed for the self-encoding strategy used here, the other modifications can be deleted in the design of a new analogue unit. For example, using the dipeptide -
10 NHCH₂CO-SCH₂CO- (Gly-SGly) as the analogue unit would investigate the role of the extra backbone methylene group. Alternatively, use of Aaa_n-S β Ala as an analogue unit would reintroduce one sidechain of the two deleted. By designing new analogue units which
15 differ from Gly-S β Ala by a single modification, the contributions of individual modifications to the polypeptide structure may be investigated after repetition of the functional selection and decoding of the new functional and nonfunctional products.

20 To identify the individual components of the mixture of protein analogues after functional selection, a strategy has been developed in which the modified polypeptides are self-encoded. A cleavable bond was introduced into the polypeptide backbone at
25 the site of modification. In the case of the analogue unit Gly-S β Ala, the cleavable bond was a thioester which is stable to the conditions of normal handling, yet can be selectively cleaved by treatment with NH₂OH at neutral pH. Chemical cleavage of the
30 thioester bond and subsequent analysis of the resulting peptide fragments by MALDI mass

5 spectrometry, gave a series of peaks which unambiguously defined the protein components present in the array before cleavage. This encoding system is especially powerful since all the information is read out in a single step from a pool of molecules free in solution.

10 The qualitative nature of the MALDI mass spectrometric readout of signature analysis experiments may allow approximate measurements of binding affinities of individual protein analogues. One solution would be to vary the conditions of the functional selection to produce a series mass spectrometric signatures; for example, selecting the protein array against a series of affinity columns 15 with increasing ligand concentrations. By monitoring the presence or absence of an individual mass spectrometric signal over a range of concentrations, an approximate K_d for the binding could be determined. Similar analyses could be performed by 20 varying other parameters such as temperature and $Gn\cdot HCl$ concentration or through an affinity elution procedure.

25 The power of a chemical synthesis approach to the study of proteins is the straightforward access to a wide range of variations in molecular structure. In this invention, backbone interactions were studied by deletion of hydrogen bonding and by insertion of an extra methylene group. However, many other types 30 of chemical modifications can be introduced using the methodology presented (supra). For example, further

- 57 -

studies could investigate the ability of the protein to tolerate restrictions in backbone conformation; aminoisobutyric acid (α,α dimethyl glycine) residues are known to restrict Ramachandran space to alpha helical conformations (Marshall et al. *Circ. Res.*, *Suppl. II* 30 and 31, 1972. 143-150) while insertion of beta turn mimics into the polypeptide chain can provide a test for such secondary structural features. The modifications possible using the tools of molecular biology are good for monitoring sidechain interactions but, with the exception of proline, do little to probe the conformational properties of the peptide backbone. The use of chemical synthesis allows for experiments which probe the tolerance of *cis* rather than *trans* peptide bond replacements and the ability of the peptide backbone to explore "D" Ramachandran space. Such experiments may give insight into the molecular characteristics of the peptide backbone. In this manner, the full range of modifications that have been used to elucidate the structure-function relationships of peptides can now be applied to proteins.

The signature analysis technique described in this example is generally applicable to proteins accessible by chemical synthesis (Muir et al. *Curr. Opin. Biotech.* 1993, 4, 420-427). With the introduction of modular chemical ligation techniques, individual domains, as well as series of these domains can be investigated in the context of larger protein molecules. Since these domains are the basic units of protein function, the systematic generation

of arrays of analogues will allow the tools of organic chemistry to be used for the elucidation of the molecular basis of protein function.

5 Example 3 (illustrates a 20 peptide example of chemical synthesis, functional separation, and readout (protein signature analysis) of self-encoded arrays of an SH3 domain with the sequence)

10 In this example, the combinatorial protein signature analysis has been applied to the N-terminal SH3 domain from c-Crk. A total of 28 chemically defined protein analogues were analysed in only two protein signature analysis experiments. Using protein signature analysis, the effect on biological function of modifying both amino acid side-chains and the 15 polypeptide backbone of the protein was determined. The latter of these, i.e. systematic backbone engineering, is unprecedented in the study of proteins. Protein signature analysis provides a framework for the systematic application of chemistry to deciphering how proteins work, and thus 20 complements the analogous chemical approaches already available for the study of nucleic acids (Min et al. (1996) *J. Am. Chem. Soc.* **118**, 6116-6120). Consequently, fundamental biological processes such 25 as protein folding, binding, and catalysis come under the scrutiny of synthetic organic chemistry.

30 In the form shown in Figure 20, protein signature analysis is a particularly useful way of looking at the chemical basis of ligand binding activity, from the viewpoint of the protein molecule.

- 59 -

As an example of this, we have applied protein signature analysis to one of the Src Homology 3 (SH3) binding domains commonly found in proteins involved in intracellular signal transduction.

5 SH3 domains are small protein modules: polypeptide chains of about 60 amino acid residues that fold to form a unique three dimensional structure, even outside the context of the longer polypeptide chain in the parent protein. It is now
10 well established that SH3 domains mediate protein-protein interactions through the recognition of short proline-rich sequences.

15 Our goal was to investigate the chemical basis of the interaction between the N-terminal SH3 domain from the cellular adaptor protein, c-Crk (residues 134-191 of the murine sequence), and its target ligand, a proline-rich peptide from the guanine nucleotide exchange protein, C3G. We wanted to change the chemical structure of the SH3 polypeptide chain
20 and observe how this affected the functional properties of the domain.

25 We decided to introduce a dramatic perturbation in the chemical structure of the protein molecule: deletion of the side chains of two adjacent amino acids in concert with the introduction of an extra backbone methylene. A thioester bond was also introduced to facilitate the identification of protein analogues (see below). The resulting Gly-[COS]- β Ala dipeptide analogue unit can be compared
30 with a native dipeptide sequence (see Figure 22A).

- 60 -

A Synthesis of an array of SH3 analogues

We initially focused on a sequence of twenty amino acids near the middle of the SH3 polypeptide chain, the region c-Crk(146-165). As a first step an array of nineteen protein analogues was chemically synthesized by placing the Gly-[COS]- β Ala dipeptide unit at each possible dipeptide position within the twenty amino acid stretch. In this synthesis, we used the modified stepwise solid phase peptide synthesis (SPPS) approach that is described in examples 1 and 2 (supra). This method made it possible to prepare all members of this array of analogues simultaneously in the course of a single synthesis (see Figure 1). This modified split-resin procedure ensured that each individual polypeptide chain in the final product mixture contained only one analogue unit at a single defined position. Stepwise synthesis of the full-length 58 residue SH3 domain polypeptide, with the introduction of a chemical perturbation at nineteen defined positions of the polypeptide chain according to such a split-resin process, gave an array of analogues as a single product mixture containing the nineteen desired molecular species.

B. Functional selection by affinity chromatography

The next task was to subject these synthetic products to functional selection. The N-terminal SH3 domain from c-Crk specifically recognizes a ten residue proline-rich sequence from the protein, C3G.

- 61 -

A synthetic peptide containing the proline-rich C3G sequence was covalently immobilized on commercially available derivatized agarose beads. In preliminary experiments, a synthetic SH3 domain corresponding to the wild-type sequence was found to bind specifically to the C3G peptide affinity column; prolonged washing with high salt buffer did not elute the synthetic SH3 protein, whereas the use of stronger conditions that disrupt the specific interactions, in this case 6M guanidine.HCl, led to elution of the protein with ~85% recovery of the applied material. No specific binding of the synthetic c-Crk SH3 domain to a control affinity column containing leucine enkephalin was observed. These procedures have previously been described in example 2 (supra).

Having established the validity of the affinity column assay, the effects of the dipeptide analogue units on the binding properties of the SH3 domain were evaluated. The nineteen member array of synthetic analogues of the 58 residue domain was passed as one pool over the C3G peptide affinity column, giving rise to binding and non-binding populations.

25 C. Readout of self-encoded arrays of protein analogues

The third and final step was then to determine which protein analogues were present in each of the two pools. The identification of individual molecular species in a pool of closely related protein analogues is a formidable analytical challenge. One

- 62 -

way to determine the molecular composition of such a mixture of protein analogues is to combine mass spectrometry with the synthetic chemistry approach, as schematically illustrated in Figure 24.

5 The readout of all members of each pool of protein analogues was accomplished in a single step using a chemical decoding approach, similar in concept to that already described for use with nucleic acid libraries (supra). A latent readout
10 chemistry was built into each molecule in the course of the preparation of the protein array by total chemical synthesis.

15 The analogue unit contained a unique thioester chemical cleavage site which allowed us to chemoselectively 'unzip' (see Figure 22B) the mixture of analogue polypeptide chains found in a particular pool, binding or non-binding. When examined by matrix assisted laser desorption ionization (MALDI) mass spectrometry (Chait, B.T. & Kent, S.B.H. (1992).
20 Weighing naked proteins: Practical high accuracy mass measurement of peptides and proteins. *Science* 257, 1885-1894), the resulting sets of 'decoded' peptide fragments gave characteristic signatures that could be interpreted as follows. Each component of the mass
25 spectrometric signature reflected the presence of the corresponding full-length polypeptide chain (containing the analogue unit) in that pool of intact protein analogues. Furthermore, the position of the analogue unit in the original 58 residue c-Crk SH3
30 protein analogue was defined by the position of the

- 63 -

corresponding signal in the mass spectrometric signature, as schematically illustrated in Figure 24 (Chait et al. (1993) *Science* **262**, 89-92; Zhao et al. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 4020-402).

5 E. Role of SH3 Backbone

The three components of protein signature analysis - synthesis, selection, and readout - have previously been described in detail in a series of model studies (examples 1 and 2 *supra*). Here we have applied them to an SH3 domain in order to elucidate the chemical basis of ligand binding. The results obtained from applying functional selection/chemical readout to the 19-member array of protein analogues corresponding to the N-terminal SH3 domain from c-Crk are shown in Figure 25. The mixture of synthetic protein analogues was passed over a C3G-peptide affinity column, to assay for binding activity. The signature of the parent array of protein analogues (Figure 25B) is compared with the signature of the pool that showed binding activity (Figure 25C).

Eight (out of nineteen) members of the array of protein analogues bound to the C3G peptide. Perhaps of most interest was the pattern of binding and non-binding observed for proteins modified within the c-Crk(146-152) region. This sequence of the SH3 protein corresponds to the so called 'RT loop', a region known to be involved in ligand binding throughout the SH3 domain family. The sequence of this part of the c-Crk SH3 polypeptide is: [-Asn¹⁴⁶-Asp-Glu-Glu-Asp-Leu-

- 64 -

Pro¹⁵²-]. It is evident from the signature of the functional pool of SH3 analogues (Figure 25C) that binding to the C3G-derived peptide ligand occurred even when the side chains of the Asp¹⁴⁷ or Glu¹⁴⁹ residues in the SH3 domain had been removed. Thus, interaction with the Asp¹⁴⁷ or Glu¹⁴⁹ side chain carboxyls was not essential for binding (note that the effect of the replacement of these two residues on binding specificity cannot be inferred from this experiment). In contrast, removal of the side chain of Asp¹⁵⁰ by substitution with either the Gly or β Ala portion of the dipeptide analogue unit virtually eliminated binding; restoration of Asp¹⁵⁰ restored binding. It should be noted that Asp¹⁴⁷ and Asp¹⁵⁰ are both conserved in the viral form of the protein, v-Crk, whereas Glu¹⁴⁹ is replaced with a glycine residue (Mayer et al. (1993) *J. Virol.* **64**, 3581-3589).

These data are intriguing and offer experimental support for a difference in the roles of the three acidic side chains in ligand binding, as previously suggested by the X-ray crystallographic data (Wu, X., et al. (1995) *Structure* **3**, 215-226.). As shown in Figure 31, all three of the side chain carboxylate functionalities in residues Asp¹⁴⁷, Glu¹⁴⁹, and Asp¹⁵⁰ of the c-Crk SH3 domain make specific interactions with the side chain -"NH₂" of Lys^{*} in the ligand. From the protein signature analysis results presented here we can infer that the primary determinant of binding in this region of the SH3 molecule is the Asp¹⁵⁰ side chain carboxylate.

- 65 -

The interaction of Asp¹⁴⁷ and Glu¹⁴⁹ side chains with the ligand peptide may play a different role, perhaps affecting the specificity of binding by discriminating between Lys and Arg side chains at 5 this position. The predominant role of Asp¹⁵⁰ and the different roles of Asp¹⁴⁷ and Glu¹⁴⁹ were both suggested by the crystallography data. In the crystal structure, the -εNH₂- group of the lysine residue (in the Pro-rich peptide ligand) forms a hydrogen bond to 10 an oxygen atom in the side-chain carboxylate of Asp¹⁵⁰, using the preferred *syn* orientation of the oxygen lone electron pair as shown (Figure 31), whereas the hydrogen bonds to Asp¹⁴⁷ and Glu¹⁴⁹ are in the less favoured *anti* orientation.

15 The signature analysis data shown in Figure 25C are consistent with this crystallographic observation, because replacement of Asp¹⁵⁰ resulted in gross loss of binding activity, whereas replacement of Asp¹⁴⁷ or Glu¹⁴⁹ did not.

20 Productive application of the signature analysis approach does not require knowledge of the three dimensional structure of a protein domain. However, the three dimensional structure can be used together with protein signature analysis to give additional 25 insights into the chemical basis of protein function. In the example given here, the combination of signature analysis data with structural studies gave a more informative interpretation of the molecular basis of ligand binding than would have been possible 30 with protein signature data alone. These results also

- 66 -

show the potential of the protein signature analysis technique to illuminate the chemical reality of mechanisms suggested by the structural data.

F. Role of SH3 backbone

5 As shown in Figure 25, eight of the 19 protein analogues displayed binding activity, while eleven were inactive. We have discussed the implications of these observations for the roles of specific amino acid side chains (above). How can we relate this data
10 to other aspects of the chemical basis of the binding function of the SH3 domain? The non-functional members of the array of protein analogues could owe their inactivity to any or all of the following factors: deletion of amino acid side chains;
15 insertion of an extra methylene in the polypeptide backbone; or, deletion of the H-bonding ability of the central amide moiety in the analogue structure. Information bearing on these possibilities can be simply obtained by making another array containing
20 alternative analogue structures covering the region of interest in the SH3 domain.

In this case, we made a second nine-membered array to the region cCrk(156-165) using -Gly-[COS]-Gly- as an analogue unit in which the additional methylene of the original analogue unit was not present (Figure 22A):

The signature obtained after functional separation, based on binding to the Pro-rich C3G peptide affinity column, and readout of this new

- 67 -

nine-membered array of protein analogues is shown in Figure 27. The signature in Figure 27B represents an expansion of the signature data shown in Figure 25C, but focussing on the region corresponding to
5 replacement of residues 156-165 of the SH3 domain by the Gly-[COS]- β Ala analogue. Of the nine Gly-[COS]- β Ala-containing SH3 analogues in this region, only four bound to the C3G peptide affinity column. The other five analogues did not bind under the
10 conditions used. This pattern is identical to that observed for this region in a data set in which only these nine analogues were analysed (example 2; *supra*).

By contrast, all nine (-Gly-[COS]-Gly-) containing protein analogues exhibited appreciable
15 binding activity in an identical assay (Figure 27A). That so many of the protein analogues retained specific binding activity is a remarkable result, given the very substantial nature of the chemical changes made in the polypeptide chain. The data show
20 that neither the pairwise deletion of side chains nor deletion of the H-bond in the central amide moiety of the analogue structure was responsible for the lack of binding exhibited by the five inactive members of the original -Gly-[COS]- β Ala-containing array of
25 protein analogues covering this region. Rather, it can be inferred from comparison of the two sets of data that the observed lack of binding activity was caused by insertion of the extra methylene in the polypeptide backbone by the original analogue unit.
30 Thus, it appears that the region defined by residues c-Crk(156-161) is less tolerant to backbone

- 68 -

engineering than the region defined by residues c-Crk (161-165) (Figure 32). The affinity binding assay used does not discriminate between a gross structural perturbation and a purely functional effect, since 5 both could result in a loss of activity. However, none of the amino acids in the region being studied (residues 156-165) interact directly with the ligand, suggesting that the observed effects may be structural in origin.

10

MATERIALS AND METHODS

General

Analytical HPLC was performed on a Hewlett-Packard 1050 system with 214 nm detection using a Vydac C18 column (5 μ m, 4.6 x 150 mm) at a flow rate of 1 mL/min. All runs used a linear 0%-67% B gradient where buffer A was 0.1% TFA in H₂O and buffer B was 90% acetonitrile, 10% H₂O, 0.09% TFA. Electrospray mass spectrometric analysis of all synthetic peptides was performed on a Sciex API-III triple quadrupole electrospray mass spectrometer. Calculated masses were obtained using the program MacProMass (Sunil Vemuri and Terry Lee, City of Hope, Duarte, CA); Buffer B was 90% acetonitrile, 10% H₂O, 0.09% TFA. Semipreparative HPLC was performed on a Rainin HPXL dual pump system using a Vydac C18 column (10 μ m, 10 x 250 mm) at 3 mL/min with detection on a Dynamax UV detector.

20 **Matrix-Assisted Laser Desorption Ionization Mass Spectrometry (MALDI) (example 1):**

Mass spectra were recorded using a Vestec Model VT 2000 laser desorption, linear time-of-flight mass spectrometer. Samples were desorbed/ionized using the focused output of a 355 nm frequency tripled Lumonics Model HY 400 Nd:YAG laser (Lumonics, Kanata, ON, Canada). Ions were accelerated through a dual-stage source to a total potential of 30 keV and detected by a 20-stage focused mesh. All spectra were acquired in the positive ion mode and summed over 50

- 70 -

laser pulses. Time-to-mass conversion was accomplished by internal calibration using the $[M+H]^+$ and $[M+2H]^{2+}$ ion signals from a standard peptide (MW 2419.1 Da). Samples were prepared by dissolving the 5 crude peptide array in 1:1 acetonitrile: H_2O , 0.1% TFA to a concentration of 1-10 μM per peptide component. 2 μL of this solution was mixed with 5 μL of a saturated solution of 2,5-dihydroxybenzoic acid (DHB) in the same solvent. Ultimately, 2 μL of this mixture 10 containing ~1-10 pmoles of each peptide component was added to a stainless steel probe tip (3.14 mm^2 .) and the solvent allowed to evaporate under ambient conditions.

Solid Phase Peptide Synthesis (example 1):

15 Except where noted, all peptides were synthesized manually according to the *in situ* neutralization/HBTU activation protocol for Boc solid phase synthesis as previously described (Schnölzer et al. *Int. J. Pept. Protein Res.* 1992, 40, 180-193). 20 The peptides were synthesized on (4-Me)benzhydrylamine-copoly(styrene-1% DVB)-resin (Peninsula Laboratories, 0.93 mmol/g) which after HF cleavage gives the C-terminal amide. The peptides were deprotected and cleaved from the resin by 25 treatment with 10 mL HF, containing 5% anisole, for one hour at 0°C. After evaporation of the HF, the crude peptide product was precipitated and washed with diethyl ether, dissolved in 1:1 Acetonitrile: H_2O containing 0.1% TFA, and lyophilized.

- 71 -

Synthesis of Boc-Gly-SCH₂CH₂COOH (example 1):

Synthesis of Boc-Gly-SCH₂CH₂COOH was based on a previously published procedure (Hojo et al. *Bull. Chem. Soc. Jpn.* 1991, 64, 111-117). To a solution of Boc-Gly-OSuc (1.36 g, 5 mmol; Sigma) dissolved in 50 mL CH₂Cl₂, 3-mercaptopropionic acid (0.5 g, 5 mmol; Aldrich) and *N,N*-diisopropylethylamine (DIEA; Sigma) (1.0 g, 7.5 mmol) were added, and the resulting solution was stirred at room temperature for 15 hours. The solvent was reduced by evaporation under reduced pressure and the resulting oil was dissolved in ethyl acetate. After two washes with 0.1 M HCl and four washes with saturated aqueous NaCl, the ethyl acetate layer was dried over magnesium sulfate. Following concentration, the resulting oil was dissolved in 40 mL diethyl ether. Dicyclohexylamine (DCHA) (4.5 mmol, 1 g) was added dropwise, giving crystals which were recrystallized from hot ethyl acetate. The DCHA salt was suspended in ethyl acetate and extracted with 0.05 M citric acid. After three washes with saturated NaCl, the ethylacetate layer was dried over magnesium sulfate. After the solution was filtered and concentrated, trituration with hexane gave Boc-Gly-SCH₂CH₂COOH as a solid (750 mg, 62 %) FAB MS for C₁₀H₁₆O₅N₁S₁Na₁, MW obsv: 286.0728 Da, calc: 286.0725; melting point 103°-105° C (104°-106°).

Synthesis of the nine component peptide analogue array of the parent sequence PFKK-[GDILRIRDKP]-EEAcpRLKLKAR (example 1):

- 72 -

The sequence EEAcRLKLKAR was synthesized in reaction vessel A, Figure 4, on 0.1 mmol MBHA resin. Onto this sequence, an array of nine Gly-S β Ala substituted peptide analogues was synthesized through 5 the sequence GDILRIRDKP using the protocol described below. Boc-Gly-SCH₂CH₂COOH (0.25 mmol, 66 mg) was preactivated for one hour with DIC (0.25 mmol, 39 μ L; Aldrich) and HOBt (0.25 mmol, 34 mg; Aldrich) in 600 μ L DMF (~0.4 M; Aldrich), and used for five 10 consecutive cycles (125 μ L/cycle), after which a second 0.25 mmol of the dipeptide analogue was activated under the same conditions and used for the remaining four cycles.

Cycle 1:

15 First, the N^α-Boc deprotected peptide-resin (0.10 mmol) was suspended in 10 mL DMF. One milliliter (~0.01 mmol) of the suspension was removed and added to a small fritted funnel. This sample was then neutralized for 1 min with 10% DIEA in DMF, drained, 20 placed in position 1 and reacted with the activated thioester dipeptide analogue (125 μ L, 0.05 mmol). During neutralization of the sample, the first subsequent activated amino acid, Boc-proline, was coupled to the resin in vessel A using manual *in situ* 25 neutralization synthetic cycles with HBTU as the activating agent. After 20 minutes coupling, the peptide-resin in vessel A was washed with DMF, treated with TFA and washed again with DMF. The first peptide-resin sample was then moved to position 30 1' where dipeptide coupling was allowed to continue.

- 73 -

Cycle 2:

The deprotected peptide-resin in vessel A was suspended in 9 mL DMF and 1 mL (~0.01 mmol) was transferred to a second small fritted funnel and 5 placed in position 1. After neutralization of this sample, 125 μ L (0.05 mmol) of the activated dipeptide was added to the sample which was placed in the now open position 1. At the same time, activated Boc-Lys was added to Vessel A. After the lysine coupling in 10 vessel A was complete, the first (dipeptide analogue) peptide-resin sample was transferred from position 1' to reaction vessel B. The peptide-resins in A and B were then deprotected with TFA, washed and finally the sample in position 1 was moved to position 1'.

15 This procedure was continued for a total of 10 cycles of chain elongation, with the final cycle skipping the removal of resin from vessel A. Finally, the sequence PFKK was added to the peptide-resin in vessel B, by stepwise SPPS. Following 20 deprotection and cleavage from the resin, the lyophilized peptide analogue array was analyzed by analytical reverse phase HPLC, electrospray mass spectrometry, and MALDI mass spectrometry.

Synthesis of the Thioester-Containing Model Peptide:
25 **LYRA-Gly-S β Ala-YGGFL-amide (example 1):**

YGGFL-4-MeBHA-resin was synthesized on an 0.04 mmol scale using standard manual Boc chemistry protocols. Boc-Gly-S β Ala (0.2 mmol, 53 mg) and HOBt

- 74 -

(0.2 mmol, 28 mg) were dissolved in 1mL DMF to which DIC (0.2 mmol, 33 μ L) was added. After 30 min, the activated dipeptide was added to the deprotected, neutralized (10% DIEA in DMF, 1 min) YGGFL-MBHA resin and allowed to couple for one hour. The sequence was completed following standard manual cycles to synthesize LYRA-Gly-S β Ala-YGGFL-4-MeBHA resin. Following deprotection and cleavage from the resin, the lyophilized peptide was characterized by analytical HPLC and by electrospray mass spectrometry. Observed mass: 1203.5 \pm 0.4 Da, calculated mass (average isotope composition): 1203.4 Da.

**Cleavage of Thioester-Containing Model Peptide
(example 1):**

The peptide LYRA-Gly-S β Ala-YGGFL·amide, was dissolved in 200 μ L of 100 mM Tris pH 9.0, 1 M Gn·HCl, vortexed vigorously for 10 seconds and left at 23°C for 30 minutes. No hydrolysis was observed under these conditions. However, addition of 20 μ L of 1 M NaOH (to pH~13) gave complete hydrolysis after just 10 minutes. Another sample of the thioester-containing peptide was dissolved in 1 M NH₂OH, 200 mM NH₄HCO₃, pH 6.0 for 30 minutes and completely cleaved into LYRAG-NHOH (observed mass: 593.5 \pm 0.5 Da, calculated mass (average isotope composition): 593.7 Da) and SHCH₂CH₂CO-YGGFL·amide (observed mass 642.5 \pm 0.5 Da, calculated mass (average isotope composition): 642.8 Da).

- 75 -

Synthesis of Ester-Containing Model Peptide: YKLFAla-[COO]-LeuYGGFL'amide (example 1):

The ester-containing model peptide was synthesized by a previously established procedure (Bramson et al. *J. Biol. Chem.* 1985, 260, 15452-15457). 5 2-Hydroxyisocaproic acid ('Leuceic acid') (1.0 mmol, 131 mg) and HOBT (1.1 mmol, 150 mg) were cooled to 0°C in 2 mL 1:1 DMF/CH₂Cl₂, activated with DIC (1.0 mmol) for 15 min and added to NH₂-YGGFL-4-MeBHA resin 10 (0.1 mmol). N-ethylmorpholine (0.25 mmol) was added and coupling proceeded for 30 min. The ester bond was created by activating Boc-Ala (1.0 mmol, 190 mg) with 4-dimethylaminopyridine (0.05 mmol,), DIC (1.0 mmol) and N-ethylmorpholine (0.25 mmol) and reacting 15 with the (α -hydroxy)acyl peptide-resin for 2 hours. The peptide chain assembly was completed by manual stepwise SPPS using *in situ* neutralization protocols. The peptide was then deprotected and cleaved from the resin and analyzed by analytical HPLC and 20 characterized by electrospray mass spectrometry; [observed mass: 1291.0 \pm 0.5 Da, calculated mass (average isotope composition): 1290.6 Da].

Cleavage of the Ester-Containing Model Peptide (example 1):

25 The ester-containing peptide YKLFAla-[COO]-LeuYGGFL'amide was allowed to stand in 6M Guanidine HCl, 100 mM Na phosphate pH 10 for six hours. Complete hydrolysis was observed. The resulting peptides were separated by analytical HPLC and

characterized by electrospray mass spectrometry. (YKLFA-OH, [observed mass: 640.5±0.5 Da, calculated mass (average isotope composition): 640.8 Da]; HO-LYGGFL-amide, [observed mass: 669.0±0.5 Da, calculated mass (average isotope composition): 668.8 Da]. By contrast, the ester-containing peptide was completely stable to treatment with 1 M NH₂OH, 200 mM NH₄HCO₃, pH 6.0 for up to 12 hours, as monitored by analytical HPLC.

10 **Hydroxylamine Cleavage of the Nine Component Peptide Gly-S₈Ala Analogue Array of the Parent Sequence PFKK-[GDILRIRDKP]-EEAcpRLKLKAR (example 1):**

15 0.3 mg (0.1 µmol/component) of the peptide array was dissolved in 100 µL 1M NH₂OH.HCl, 200mM NH₄HCO₃, pH 6.0. After 30 minutes, the array was analyzed by analytical reverse phase HPLC and MALDI mass spectrometry.

20 **Solid Phase Peptide Synthesis (example 2)**
Except where noted all peptides were synthesized according to the machine-assisted *in situ* neutralization/HBTU activation protocol for Boc-solid phase chemistry as previously described (Schnölzer et al. *Int. J. Pept. Protein Res.* 1992, 40, 180-193) using a modified Applied Biosystems 430A peptide synthesizer. Following synthesis, the N^oBoc group was removed, and the peptide cleaved from the resin with simultaneous removal of sidechain protecting groups by treatment for 1 hour at 0°C with anhydrous HF containing 5% anisole or 5% p-cresol as a scavenger.

- 77 -

After evaporation of the HF, the crude peptide was precipitated and washed with cold diethyl ether, dissolved in 1:1 acetonitrile:H₂O containing 0.1% TFA, filtered to remove the resin, and lyophilized.

5 **Synthesis of native N-terminal cCrk SH3 domain (cCrk residues 134-191) (example 2):**

The 58 amino acid residue polypeptide was synthesized using 0.12 mmol Boc-Arg(Tos)-OCH₂-Pam resin, loading 0.59 mmol/g (Applied Biosystems, Foster City, CA). Standard sidechain protecting groups (Schnölzer et al. *Int. J. Pept. Protein Res.* 1992, 40, 180-193) were used except for the tryptophan indole moiety which was left unprotected because subsequent syntheses of base labile analogues would not permit the nucleophilic removal of the usual formyl protecting group. HF cleavage of 300 mg of peptide resin from the resin gave 195 mg of lyophilized crude peptide. A 5 mg sample of the crude peptide was purified by semipreparative HPLC on a 20%-40%B gradient over 45 minutes to give 1.4 mg purified product (23% yield, calculated from the original loading of the resin). The purified peptide product was a single peak by analytical HPLC and was pure by electrospray mass spectrometry: Observed mass 6962±1 Da: Calculated mass for C₃₁H₄₇N₈O₉S₁ 6961.8 Da (average isotope distribution).

Functional characterization of the synthetic 58-residue cCrk N-terminal SH3 domain:

- 78 -

The affinity of the synthetic SH3 domain to two C3G derived peptides was determined by measuring the increase in the protein domain tryptophan fluorescence upon ligand binding, following the 5 procedure described in Lim et al. *Protein Science*. 1994, 3, 1261-1266. The purified and lyophilized 58 residue polypeptide chain (0.2 mg) was folded by dissolving in 0.6 mL of 20 mM Hepes, 60 mM NaCl, pH 7.3 to produce a folded protein solution (~50 μ M). 10 Peptide stock solutions of both the C3G-derived peptide [PPPALPPKKR.amide] (71.9 μ M) and the C3G-derived peptide designed for attachment to an affinity column [Ac-CWAcP-PPPALPPKKR.amide] (71.5 μ M) were obtained by dissolving the lyophilized peptides 15 in the same buffer. Peptide concentrations were determined by quantitative amino acid analysis.

Synthesis of the nine component analogue array of cCrk (134-191) (example 2):

The target array of analogue polypeptide chains 20 is shown in **Figure 23**. The first part of the sequence, corresponding to cCrk (166-191), was synthesized on a 0.2 mmol Boc-Arg(Tos)-OCH₂-Pam resin, (0.59 mmol/g) using machine-assisted synthetic cycles. The array of polypeptide analogues was 25 manually synthesized on ~0.05 mmol (250 mg) of this peptide resin by a modified split-resin procedure previously described in Hayashibara et al. *J. Am. Chem. Soc.* 1991, 113, 5103-5106 (similar readout system developed for use in DNA systems). Boc-Gly- 30 S β Ala-OH (0.25 mmol, 66 mg) was preactivated for one

- 79 -

hour with DIC (0.25 mmol, 38 μ L; Aldrich) and HOBr (0.25 mmol; Aldrich) in DMF: total volume 650 μ L. The apparatus for manual synthesis consisted of two standard manual synthesis reaction vessels, labeled A and B and two small fritted funnels in a test tube rack in positions 1 and 1'.

5 **Cycle 1:**

First, the N^o-Boc-deprotected cCrk(166-191)peptide-resin (50 μ mol) was suspended in 10 mL DMF. One milliliter (~5 μ mol) of the suspension was removed and added to a small fritted funnel. This sample was then neutralized for 1 min with 10% DIEA in DMF, drained, placed in position 1 and reacted with the activated dipeptide analogue (65 μ L, 25 μ mol). During neutralization of the sample, the coupling of the first activated amino acid, Boc-Pro¹⁶⁵, to the peptide-resin in reaction vessel A was initiated using manual *in situ* neutralization synthetic cycles with HBTU as the activating agent.

10 15 20 25 After 20 minutes, vessel A was washed with DMF, treated with TFA and washed again with DMF. The removed peptide-resin sample in position 1 was then moved to position 1' where coupling of the dipeptide analogue was continued.

25 **Cycle 2:**

The peptide-resin in vessel A was suspended in 9 mL DMF and 1 mL (~5 μ mol) was transferred to a second small fritted funnel and placed in position 1. After

- 80 -

neutralization of this sample and addition of activated Lys¹⁶ to **A**, the activated dipeptide (65 μ L, 25 μ mol) was added to the sample which was placed in the now open position 1. Following the Boc-Lys¹⁶ coupling in vessel **A**, the first analogue-modified peptide resin sample in position 1' was washed with DMF and then transferred to reaction vessel **B**. The peptide-resins in reaction vessels **A** and **B** were then deprotected with TFA, washed and finally, the second peptide-resin sample in position 1 was moved to position 1', where coupling of the dipeptide analogue was continued.

The procedure described above was continued for 10 cycles, through the addition of residue 156. In the final cycle the removal of a peptide-resin sample from vessel **A** and dipeptide coupling steps were omitted. Finally, half of the mixture of peptide-resins in vessel **B** (25 μ mol total) containing nine peptide analogues, was removed and transferred to an Applied Biosystems 430A peptide synthesizer for addition of the remaining amino acids, cCrk 134-155. This procedure gave a product mixture of 58 residue peptide-resins. Following deprotection and cleavage from the resin, the lyophilized peptide array was analyzed by analytical reverse phase HPLC and MALDI mass spectrometry.

Synthesis of C3G-derived ligand and control peptides (example 2):

The peptides corresponding to the C3G-derived

- 81 -

ligand, Ac-CWAcP-PPPAlPPKKR·amide and the control, Ac-
CAcP YGGFL·amide, were synthesized on 4-methyl
benzhydrylamine resin (0.93 mmol/g Peninsula
Laboratories) and were cleaved from the resin support
5 using p-cresol as a scavenger, and then purified by
semipreparative HPLC using a 25%-50% acetonitrile
gradient over 30 minutes. The products were
characterized by ESMS. Ac-CWAcP-PPPAlPPKKR·amide;
Observed mass: 1543±1 Da. Calculated mass for
10 C₁₄H₂₁N₂O₄S₁ (average isotope composition): 1543.9 Da,
Ac-CAcP YGGFL·amide; Observed mass: 813.5±0.5 Da,
Calculated mass for C₁₀H₁₆N₂O₂S₁ (average isotope
composition): 814.0 Da.

Preparation of affinity columns (example 2):

15 The C3G-derived synthetic peptide affinity
column was prepared by adding 10 mg of lyophilized
Ac-CWAcP-PPPAlPPKKR·amide (Acp = ε-amino caproic acid)
in 2 mL of 50 mM Tris, 5 mM EDTA, pH 8.0 buffer to
Sulfolink™ resin (Pierce), equilibrated in the same
20 buffer, for 1 hour while shaking. Unreacted iodoalkyl
groups on the resin were then blocked by treatment
with 50 mM cystamine, 50 mM Tris, 5 mM EDTA, pH 8.0
buffer for 1 hour. The loading of the column was
determined by UV absorbance of the unreacted peptide
25 solution and was approximately 5 μmol/mL. A similar
procedure was used to attach the control peptide Ac-
CAcP YGGFL·amide to the another batch of the Sulfolink™
support.

Affinity selection of synthetic cCrk SH3 domain

- 82 -

(example 2) :

Lyophilized crude synthetic polypeptide corresponding to the cCrk SH3 domain (1.5 mg) and BSA (8.5 mg) were dissolved in 20 mM HEPES, 50 mM NaCl, pH 7.3 buffer (400 μ L) and loaded on to a 1 mL C3G-derived synthetic peptide affinity column pre-equilibrated with the same buffer. The column was washed with 4 mL each of 50 mM NaCl, 100 mM NaCl, 200 mM NaCl, 300 mM NaCl, 400 mM NaCl, 500 mM NaCl and 1000 mM NaCl, 0.1 M phosphate, pH 7.0 buffer. The column was then washed with 1 M NH₂OH 200 mM NH₄CO₃, pH 6.0 (4 mL) and finally eluted in 6M Gn.HCl 100 mM phosphate pH 6.5 (4 mL). Samples from all column fractions were monitored by absorbance at 280 nm and by HPLC.

Affinity chromatography the nine-membered arrays

(example 2) :

Affinity chromatography of the arrays of protein analogues was carried out as follows. The crude protein array (1.5 mg) was dissolved in 20 mM HEPES, 50 mM NaCl, pH 7.3 buffer (600 μ L) and loaded on to a 1 mL C3G-derived synthetic peptide affinity column pre-equilibrated with the same buffer. After incubation at room temperature for a period of 6-8 hours the column was washed with 0.5 M NaCl in 0.1 M sodium phosphate, pH 7.0 buffer (6 x 1 mL) to remove any non-specifically bound proteins. The first two 1 mL fractions were collected and immediately mixed with an equal volume of 1 M NH₂OH, 20 mM NH₄HCO₃, pH

- 83 -

5.5 buffer to cleave any thioester-containing protein analogues. Following a second 6 x 1 mL column wash with 1 M NaCl in 0.1 M sodium phosphate, pH 7.0 buffer, specifically-bound protein analogues were 5 chemically cleaved and eluted from the affinity column by washing with 1 M NH₂OH, 20 mM NH₄HCO₃, pH 5.5 buffer (4 x 1 mL). Samples of the eluted fractions were monitored by UV absorbance at 280 nm. This 10 procedure was used for both the C3G peptide column and for the control column loaded with Ac-C-Acp-YGGFL-amide. To accommodate MALDI analysis, both the 0.5M NaCl wash and 1 M NH₂OH fractions were desalting on a low pressure, disposable C-18 column, washed 15 with HPLC buffer A and peptide fragments eluted with 1.5 mL 60% acetonitrile in water, 0.1% TFA.

MALDI mass spectrometric analysis of peptide fragments (example 2):

20 After desalting, the affinity column fractions were analyzed by MALDI mass spectrometry. Samples were prepared by adding a 2 μ L aliquot of the 1.5 mL desalting column fraction to 5 μ L of a saturated solution of α -cyano-4-hydroxycinnamic acid in 50% acetonitrile in water, 0.1% TFA. From this mixture, 25 2 μ L, containing ~1-10 pmole of each peptide component was added to a stainless steel probe tip (3.14 mm²) and the solvent allowed to evaporate slowly under ambient conditions. Mass spectra were recorded using a prototype laser desorption, linear time-of-flight mass spectrometer from Ciphergen 30 Biosystems Inc. (Palo Alto, CA). Samples were

- 84 -

ionized using 337 nm radiation output from a nitrogen laser (Laser Science, Inc., Newton MA). All spectra were acquired in the positive ion mode and summed over 20-50 laser pulses. Time-to mass conversion was
5 accomplished by internal calibration using the $[M+H]^+$ signals from the largest and smallest peptide components in each array.

Synthesis of Peptides (example 3):

With the exception of protein arrays all
10 peptides were chemically synthesized according to optimized solid-phase methods (Schnölzer (1992) *Int. J. Pept. Protein Res.* **40**, 180-193) and purified by preparative reverse-phase HPLC using a Vydac C-18 column. In all cases, peptide composition and purity
15 were confirmed by electrospray mass spectrometry and analytical reverse-phase HPLC.

Synthesis of C-Crk SH3 Protein Arrays (example 3):

A detailed description of the split-resin procedure used (examples 1 and 2; supra). Briefly,
20 the technique involves the use of two reaction vessels with identical synthetic manipulations being carried out in each. Standard stepwise chain assembly was initiated on resin in the first vessel (0.2 mmole scale); peptide-resin samples were repeatedly removed
25 from the first vessel at each stage of the synthesis, and analogue units were introduced into the polypeptide chain by coupling as preformed HOBt esters; after modification, the samples were transferred to the second vessel for completion of

- 85 -

the chain assembly by standard stepwise chain assembly. The size of the samples was adjusted to yield approximately equal molar amounts of each protein analogue in the array (dependent on the 5 number of protein analogues in a given array). The dipeptide analogues Gly-[COS]- β Ala and Gly-[COS]-Gly were prepared as previously described (Hojo et al. (1991) *Bull. Chem. Soc. Jpn.* **64**, 111-117). Upon completion of the synthesis, each parent protein 10 array was characterized as follows: crude protein array (~1 mg) was dissolved in a cleavage buffer consisting of 1 M NH₂OH, 20 mM NH₄HCO₃, pH 6.5 buffer (1 ml) and stirred for 15 minutes. The cleaved arrays were then exchanged into a 70% CH₃CN:30% H₂O, 0.1% TFA 15 solvent system (using a 1 ml C-18 desalting column) and immediately analysed by MALDI mass spectrometry.

Synthesis of Peptide Affinity Columns (example 3):

The C3G peptide affinity column was prepared as follows. The peptide Ac-CWBPPP₁ALPPKKR.amide (B = ϵ -20 aminocaproic acid) was dissolved in 50 mM Tris, 5 mM EDTA, pH 8.0 (10 mg in 2 ml) and shaken with SulfolinkTM resin (Pierce) for 1 hour. Unreacted iodoalkyl groups on the resin were then blocked by treatment with 50 mM cystamine, 50 mM Tris, 5 mM 25 EDTA, pH 8.0 buffer. The loading of the column was determined by UV to be approximately 5 μ mole/ml. A similar procedure was used to attach the control peptide Ac-CBYGGFL.amide (YGGFL = leucine enkephalin) to the SulfolinkTM support.

Affinity Selection of Synthetic c-Crk SH3 (example 3):

HPLC purified synthetic c-Crk SH3 (1.5 mg) was dissolved in 20 mM Hepes, 50 mM NaCl, pH 7.3 buffer (400 μ l) and applied to a 1 mL C3G peptide affinity column pre-equilibrated with the same buffer. After 6-8 hours (required for optimal binding) the column was washed with, in turn: 0.5 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer (6 x 1 ml), 1 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer (6 x 1 ml) and 1 M NH₂OH, 20 mM NH₄HCO₃, pH 5.5 buffer (6 x 1 ml). The applied material did not elute from the column under any of these conditions, but was readily recovered (with ~85% yield) by washing the column with a 6 M GuHCl, 0.1 M sodium phosphate, pH 7.0 buffer (2 x 1 ml). In contrast, the synthetic SH3 domain did not specifically bind to the leucine enkaphalin control column under identical conditions to the above.

Affinity Selection of Protein Arrays (example 3):

The crude protein array (1.5 mg) was dissolved in 20 mM Hepes, 50 mM NaCl. pH 7.3 buffer (600 μ l) and loaded on to a 1 ml C3G peptide affinity column pre-equilibrated with the same buffer. After 6-8 hours the non-specifically bound material was eluted from the column by washing with 0.5 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer (6 x 1 ml). Eluted material (typically in the first and second wash) was immediately cleaved by dilution into 1 M NH₂OH, 20 mM NH₄HCO₃, pH 5.5 buffer. Following further column

- 87 -

washing with 1 M NaCl, 0.1 M sodium phosphate, pH 7.0 buffer, the specifically bound material (active pool) was chemically cleaved and simultaneously eluted from the affinity column by washing with 1 M NH₂OH, 20 mM NH₄HCO₃, pH 6.5 buffer (4 x 1 ml). Eluted fractions 5 were exchanged into a 70% CH₃CN:30% H₂O, 0.1% TFA solvent system and immediately analysed by MALDI mass spectrometry.

MALDI Analysis of Peptide Arrays (example 3):

10 All samples were prepared by adding 2 μ L of the desalted column fraction to 5 μ L of a saturated solution of α -cyano cinnaminic acid in 50 % acetonitrile in water, 0.1% TFA. From this mixture, 2 μ L, containing ~1-10 pmole of each peptide component 15 was added to a stainless steel probe tip and the solvent allowed to evaporate under ambient conditions. Mass spectra were recorded using a prototype laser desorption, linear time-of-flight mass spectrometer from Ciphergen Biosystems (Palo Alto, CA). Samples were desorbed/ionized using 337 nm 20 radiation output from a nitrogen laser (Laser Science, Inc., Newton MA). All spectra were acquired in the positive ion mode and summed over 20-50 laser pulses. Time-to mass conversion was accomplished by 25 internal calibration using the the [M+H]⁺ signals from the largest and smallest peptide components in each array.

What is claimed is:

1. A method for obtaining a molecular signature of a protein, the protein having an amino acid sequence with length m , each amino acid position within the sequence being represented by $(aa)_n$ where $1 \leq n \leq m$, the protein having a binding affinity with respect to a target molecule under binding conditions, the molecular signature being defined by a subsequence of the amino acid sequence selected from amongst positions $(aa)_n$ which, if individually replaced by a substitute amino acid, lead to a loss of binding affinity by the protein with respect to the target molecule, the method comprising the following steps:

15 Step A: providing a peptide ladder library with m peptides, each peptide being represented by $(peptide)_n$, where $1 \leq n \leq m$, each peptide having the same amino acid sequence as the protein except that position $(aa)_n$ of $(peptide)_n$ is replaced by the substitute amino acid, the substitute amino acid at position $(aa)_n$ being linked to the amino acid at position $(aa)_{n+1}$ by means of a labile bond; then

20 Step B: contacting the peptide ladder library of said step A with the target molecule under binding conditions for forming bound peptides and unbound peptides, the bound peptides being bound to the target molecule; then

25 Step C: separating the unbound peptides from the bound peptides of said Step B for obtaining

- 89 -

5 separated unbound peptides, each separated
unbound peptide having the substitute amino acid
only at positions (aa)_n corresponding to the
subsequence which defines the molecular
signature of the protein with respect to the
target molecule; then

10 Step D: cleaving the labile bond of the separated
unbound peptides obtained in said Step C for
producing peptide cleavage products, each
peptide cleave product corresponding to one of
the positions (aa)_n from the subsequence which
defines the molecular signature; then

15 Step E: detecting and identifying each of the
peptide cleavage products of said Step D; and
then

20 Step F: constructing the subsequence which defines
the molecular signature of the protein with
respect to the target molecule using the
identity of the peptide cleavage products of
said Step E.

2. A method for obtaining a molecular signature of a
protein as described in Claim 1 wherein the labile
bond within the peptide of said Step A is selected
25 from the group consisting of thioester bonds and
ester bonds.

- 90 -

3. A method for obtaining a molecular signature of a protein as described in Claim 1 wherein the substitute amino acid in said Step A is selected from the group consisting of L-alanine, L-arginine, L-aspartic acid, L-asparagine, L-cysteine, L-cystine, L-glutamic acid, L-glutamine, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan, L-tyrosine, L-valine, D-alanine, D-arginine, D-aspartic acid, D-asparagine, D-cysteine, D-cystine, D-glutamic acid, D-glutamine, D-glycine, D-histidine, D-isoleucine, D-leucine, D-lysine, D-methionine, D-phenylalanine, D-proline, D-serine, D-threonine, D-tryptophan, D-tyrosine, D-valine, L- α -aminobutyric acid, D- α -aminobutyric acid, L- γ -aminobutyric acid, D- γ -aminobutyric acid, L- ϵ -aminocaproic acid, D- ϵ -aminocaproic acid, L-homophenylalanine, D-homophenylalanine, L-alloisoleucine, D-alloisoleucine, L- β -2-napthylalanine, D- β -2-napthylalanine, L-norvaline, D-norvaline, L-ornithine, D-ornithine, L-pyridyl alanine, D-pyridyl alanine, L-2-thienylalanine, D-2-thienylalanine L-methyltyrosine, D-methyltyrosine, L-citrulline, D-citrulline, L-homocitrulline, and D-homocitrulline.

4. A method for obtaining a molecular signature of a protein as described in Claim 1 wherein, in said Step A, the substitute amino acid at position (aa)_n is attached to a second substitute amino acid to form a

- 91 -

footprint of two substitute amino acids.

5. A method for obtaining a molecular signature of a protein as described in Claim 1 wherein, in said Step 5 A, the substitute amino acid at position $(aa)_n$ is attached to a second and a third substitute amino acid to form a footprint of three substitute amino acids.

10. 6. A method for obtaining a molecular signature of a protein, the protein having an amino acid sequence with length m , each amino acid position within the sequence being represented by $(aa)_n$ where $1 \leq n \leq m$, the protein having a binding affinity with respect to a target molecule under binding conditions, the 15 molecular signature being defined by a subsequence of the amino acid sequence selected from amongst positions $(aa)_n$ which, if individually replaced by a substitute amino acid, lead to a loss of binding 20 affinity by the protein with respect to the target molecule, the method comprising the following steps:

Step A: providing a peptide ladder library with m peptides, each peptide being represented by $(peptide)_n$, where $1 \leq n \leq m$, each peptide having the same amino acid sequence as the protein except that position $(aa)_n$ of $(peptide)_n$ is replaced by the substitute amino acid, the substitute amino acid at position $(aa)_n$ being linked to the 25

- 92 -

amino acid at position $(aa)_{n+1}$ by means of a labile bond; then

5 Step B: contacting the peptide ladder library of said step A with the target molecule under binding conditions for forming bound peptides and unbound peptides, the bound peptides being bound to the target molecule; then

10 Step C: separating the unbound peptides from the bound peptides of said Step B for obtaining separated bound peptides, each separated bound peptide lacking any substitute amino acid at position $(aa)_n$ corresponding to the subsequence which defines the molecular signature of the protein with respect to the target molecule; then

15 Step D: cleaving the labile bond of the separated bound peptides obtained in said Step C for producing peptide cleavage products, each peptide cleave product corresponding to one of the positions $(aa)_n$ not included within the subsequence which defines the molecular signature; then

20 Step E: detecting and identifying each of the peptide cleavage products of said Step D; and then

25 Step F: constructing the subsequence which defines the molecuclar signature of the protein with respect to the target molecule using the

identity of the peptide cleavage products of said Step E.

7. A method for obtaining a molecular signature of a protein as described in Claim 6 wherein the labile bond within the peptide of said Step A is selected from the group consisting of thioester bonds and ester bonds.

10 8. A method for obtaining a molecular signature of a protein as described in Claim 6 wherein the substitute amino acid in said Step A is selected from the group consisting of L-alanine, L-arginine, L-aspartic acid, L-asparagine, L-cysteine, L-cystine, L-glutamic acid, L-glutamine, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan, L-tyrosine, L-valine, D-alanine, D-arginine, D-aspartic acid, D-asparagine, D-cysteine, D-cystine, D-glutamic acid, D-glutamine, D-glycine, D-histidine, D-isoleucine, D-leucine, D-lysine, D-methionine, D-phenylalanine, D-proline, D-serine, D-threonine, D-tryptophan, D-tyrosine, D-valine, L- α -aminobutyric acid, D- α -aminobutyric acid, L- γ -aminobutyric acid, D- γ -aminobutyric acid, L- ϵ -aminocaproic acid, D- ϵ -aminocaproic acid, L-homophenylalanine, D-homophenylalanine, L-alloisoleucine, D-alloisoleucine, L- β -2-naphthylalanine, D- β -2-naphthylalanine, L-norvaline, D-

15

20

25

- 94 -

norvaline, L-ornithine, D-ornithine, L-pyridyl
alanine, D-pyridyl alanine, L-2-thienylalanine, D-2-
thienylalanine L-methyltyrosine, D-methyltyrosine, L-
citrulline, D-citrulline, L-homocitrulline, and D-
5 homocitrulline.

9. A method for obtaining a molecular signature of a
protein as described in Claim 6 wherein, in said Step
A, the substitute amino acid at position $(aa)_n$ is
10 attached to a second substitute amino acid to form a
footprint of two substitute amino acids.

10. A method for obtaining a molecular signature of a
protein as described in Claim 6 wherein, in said Step
A, the substitute amino acid at position $(aa)_n$ is
15 attached to a second and a third substitute amino
acid to form a footprint of three substitute amino
acids.

20 11. A peptide ladder library corresponding to a
protein having an amino acid sequence with length m ,
each amino acid position within the sequence of the
protein being represented by $(aa)_n$ where $1 \leq n \leq m$, the
peptide ladder library comprising:

25 m peptides, each peptide being represented by
 $(peptide)_n$, where $1 \leq n \leq m$, each peptide having the same

amino acid sequence as the protein except that position (aa)_n of (peptide)_n is replaced by a substitute amino acid, the substitute amino acid at position (aa)_n being linked to the amino acid at 5 position (aa)_{n+1} by means of a labile bond.

12. A peptide ladder library as described in Claim 11 wherein the labile bond within the peptide is selected from the group consisting of thioester bonds 10 and ester bonds.

13. A peptide ladder library as described in Claim 11 wherein the substitute amino acid is selected from the group consisting of L-alanine, L-arginine, L-aspartic acid, L-asparagine, L-cysteine, L-cystine, L-glutamic acid, L-glutamine, L-glycine, L-histidine, L-isoleucine, L-leucine, L-lysine, L-methionine, L-phenylalanine, L-proline, L-serine, L-threonine, L-tryptophan, L-tyrosine, L-valine, D-alanine, D-arginine, D-aspartic acid, D-asparagine, D-cysteine, D-cystine, D-glutamic acid, D-glutamine, D-glycine, D-histidine, D-isoleucine, D-leucine, D-lysine, D-methionine, D-phenylalanine, D-proline, D-serine, D-threonine, D-tryptophan, D-tyrosine, D-valine, L- α -aminobutyric acid, D- α -aminobutyric acid, L- γ -aminobutyric acid, D- γ -aminobutyric acid, L- ϵ -aminocaproic acid, D- ϵ -aminocaproic acid, L-homophenylalanine, D-homophenylalanine, L-alloisoleucine, D-alloisoleucine, L- β -2-

- 96 -

5 naphtylalanine, D- β -2-naphtylalanine, L-norvaline, D-norvaline, L-ornithine, D-ornithine, L-pyridyl alanine, D-pyridyl alanine, L-2-thienylalanine, D-2-thienylalanine L-methyltyrosine, D-methyltyrosine, L-citrulline, D-citrulline, L-homocitrulline, and D-homocitrulline.

14. A peptide ladder library as described in Claim 11
wherein the substitute amino acid at position (aa)_n
is attached to a second substitute amino acid to form
a footprint of two substitute amino acids.

15. A peptide ladder library as described in Claim 11
wherein the substitute amino acid at position (aa)_n
is attached to a second and a third substitute amino
acid to form a footprint of three substitute amino
acids.

20 16. A method for constructing a peptide ladder library corresponding to a protein having an amino acid sequence with length m , each amino acid position of the protein being represented by $(aa)_n$ where $1 \leq n \leq m$, the peptide library including m peptides, each peptide being represented by $(peptide)_n$, where $1 \leq n \leq m$, each peptide having the same amino acid sequence as the protein except that position $(aa)_n$ of $(peptide)_n$ is replaced by a substitute amino acid, the

- 97 -

substitute amino acid at position $(aa)_n$ being linked to the amino acid at position $(aa)_{n+1}$ by means of a labile bond, the method comprising the following steps:

5 Step A: providing a first reaction vessel containing a first pool of nascent peptides having a length of $m-n$ and the amino acid sequence between $n+1$ and m of the protein, the nascent peptides being attached to a matrix material;

10 Step B: providing a second reaction vessel with a first pool of nascent ladder peptides having a length of $m-n$ and the amino acid sequence between $n+1$ and m of the protein except that each (nascent ladder peptide)_p has the substitute amino acid at position $(aa)_p$, where $n+1 \leq p \leq m$, the nascent ladder peptides being attached to a matrix material; then

15 Step C: transferring an aliquot of matrix material from the first reaction vessel to a third reaction vessel; then

20 Step D: elongating the first pool of nascent peptides in the first reaction vessel by addition of the amino acid of position $(aa)_n$ to form a second pool of nascent peptides having a length of $m-n+1$;

25 Step E: elongating the aliquot of nascent peptides in the third reaction vessel by addition of the

- 98 -

substitute amino acid of position $(aa)_n$ by means of labile bond to form a nascent ladder (peptide)_n having a length of $m-n+1$;

5 Step D: elongating the first pool of nascent peptide ladders in the third reaction vessel by addition of the amino acid of position $(aa)_n$ to form a partial second pool of nascent peptide ladders having a length of $m-n+1$; then

10 Step F: transferring the product of the third reaction vessel in said Step E to the second reaction vessel to complete the second pool of nascent peptide ladders having a length of $m-n+1$; and then

15 Step G: repeating said Steps C through F until $n=1$ and the second reaction vessel contains the peptide ladder library.

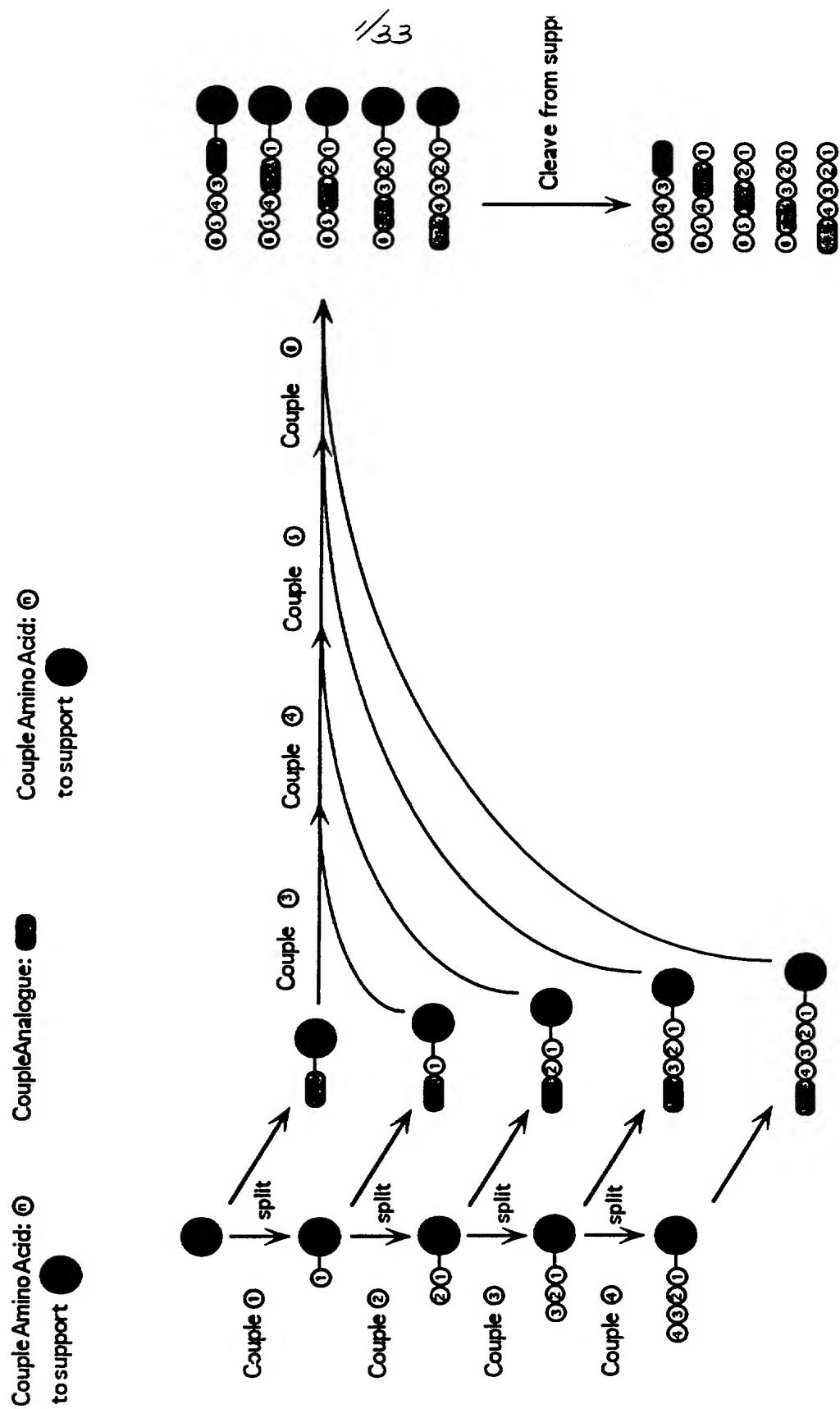
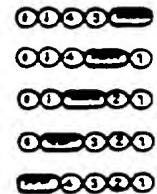


FIGURE 1

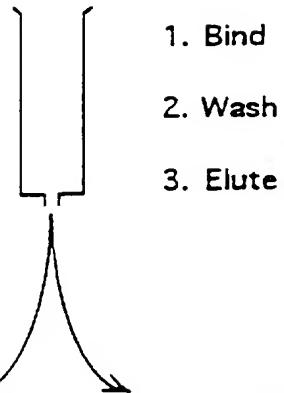
2/33

Step 1: Synthesis of polypeptide family.

Synthetic marker scanned through target polypeptide sequence. See figure 1 for details.

**Step 2: Selection of polypeptide family.**

Mixture can be screened for some structural or functional property. For example, affinity chromatography could be used to separate binding from non-binding components of the polypeptide family.

**Step 3: Chemical cleavage and analysis of fragments.**

Following selection, the synthetic marker is selectively cleaved with say a nucleophile, and the resulting peptide fragments analysed by HPLC and mass spectrometry. Once identified these fragments allow the original location of the marker within the parent molecule to be determined.

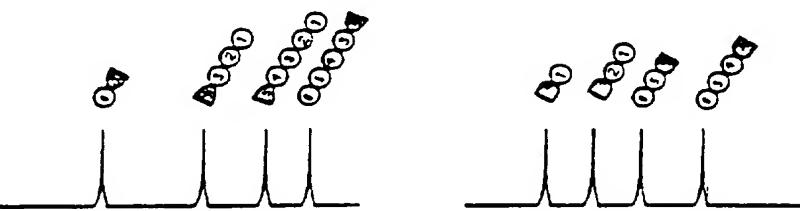


FIGURE 2

3/33

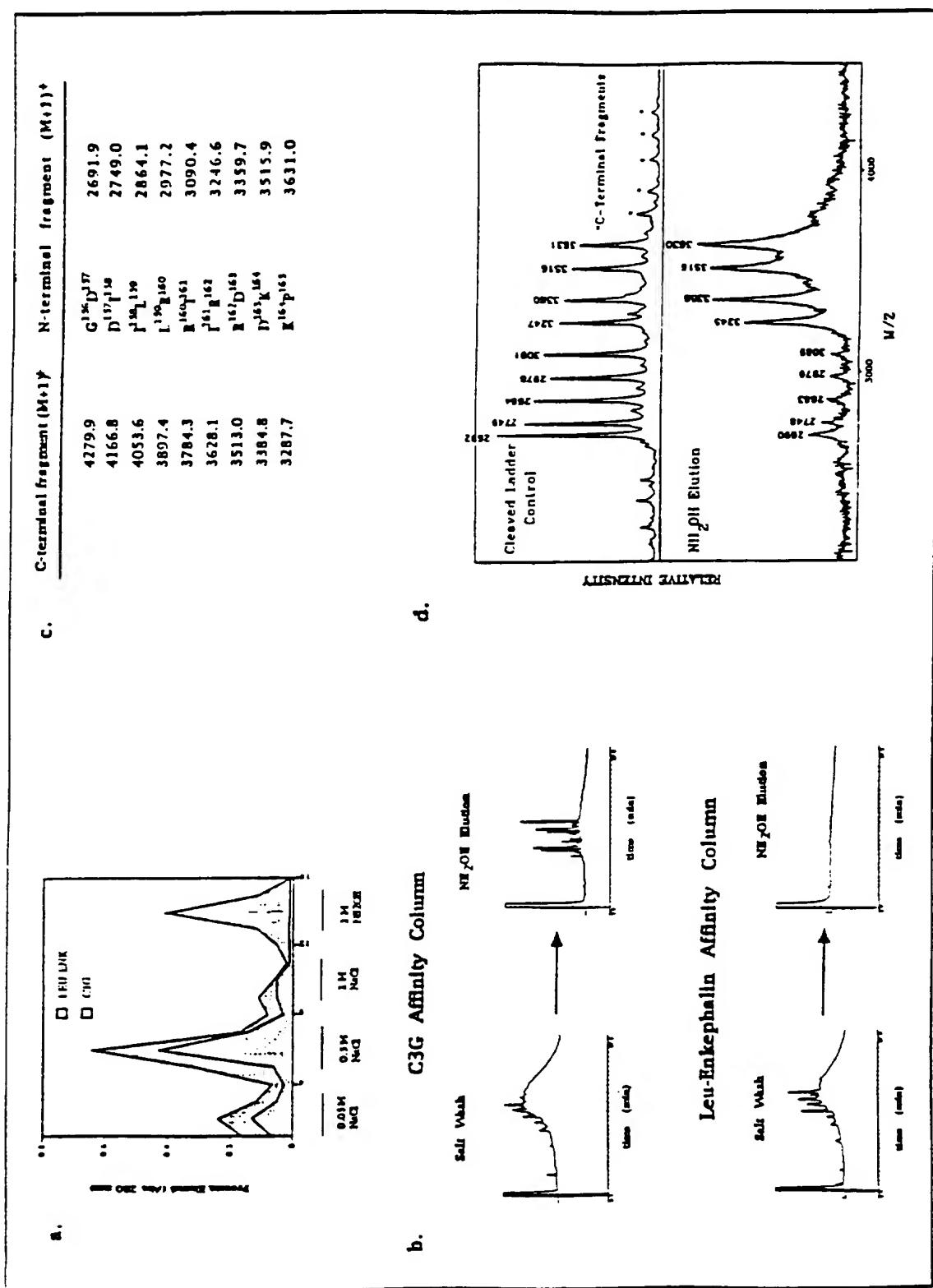
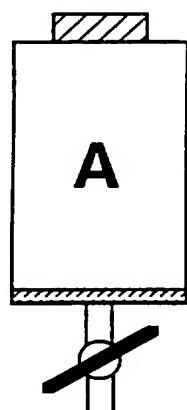


FIGURE 3

44/33

Vessel A

sample

Modify

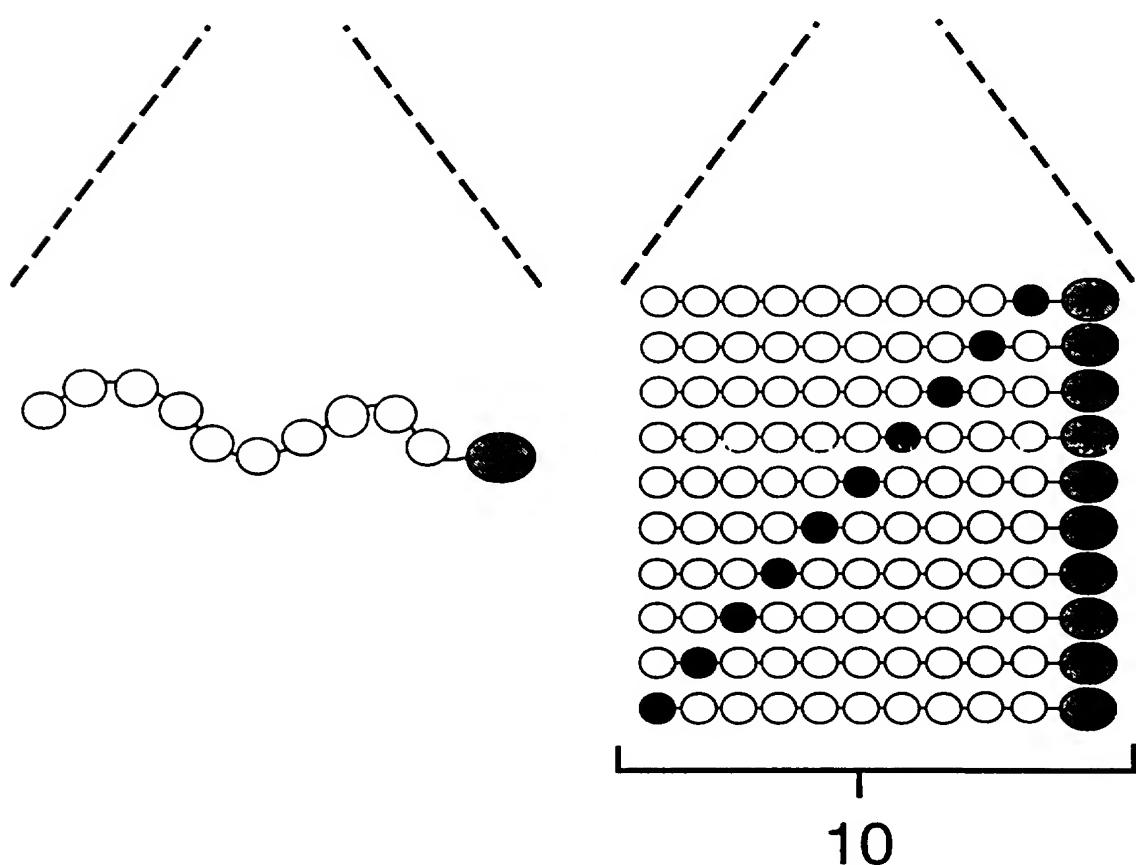
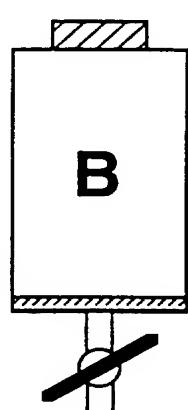
Vessel B

FIGURE 4

5/33

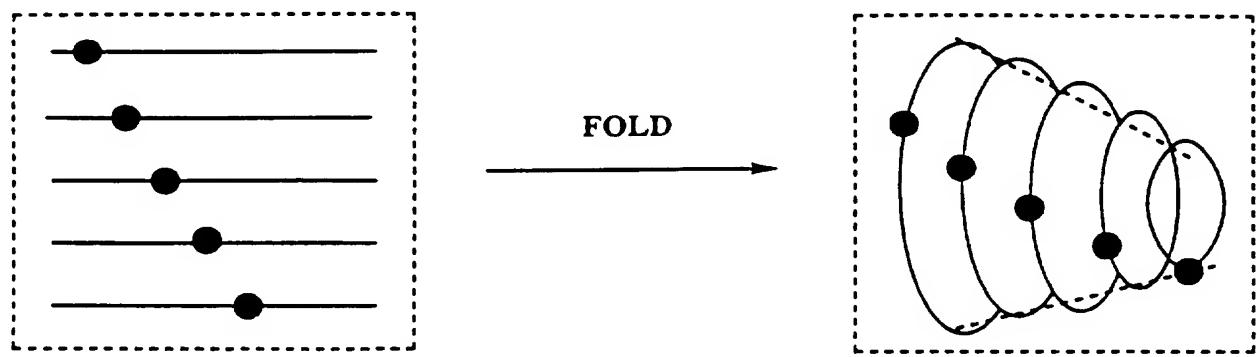


FIGURE 5

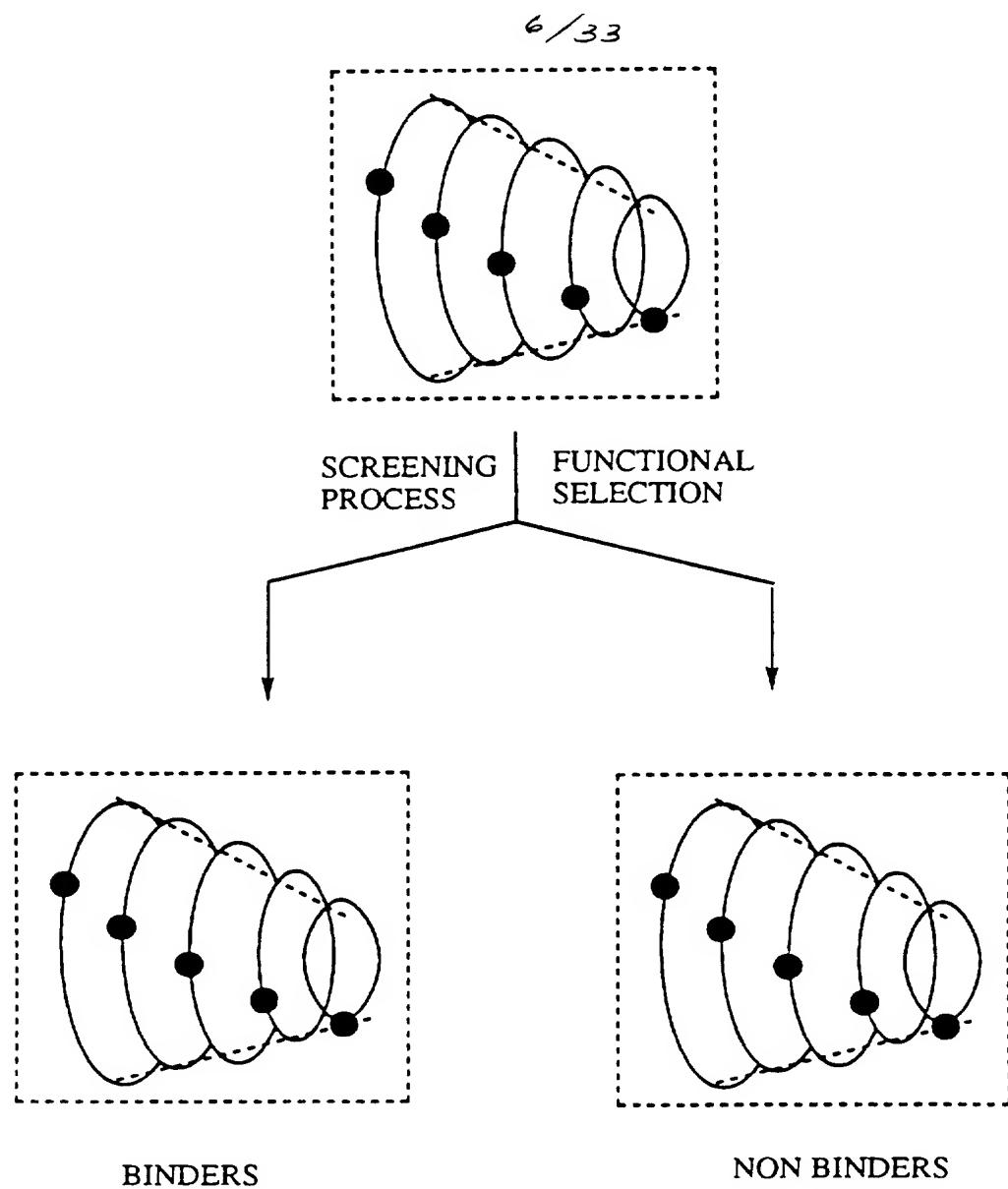
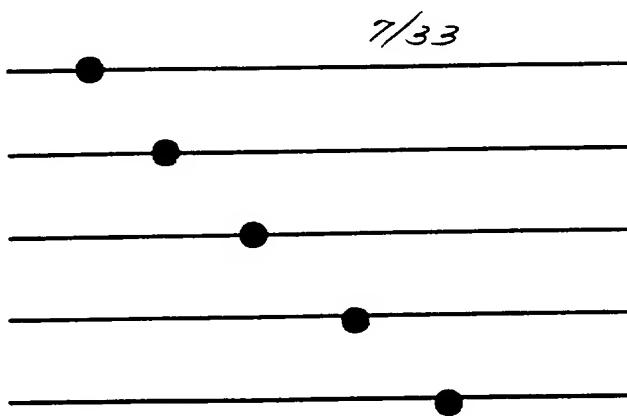


FIGURE 6



SCREENED FAMILY OF PEPTIDES EITHER BINDERS OR NONBINDERS

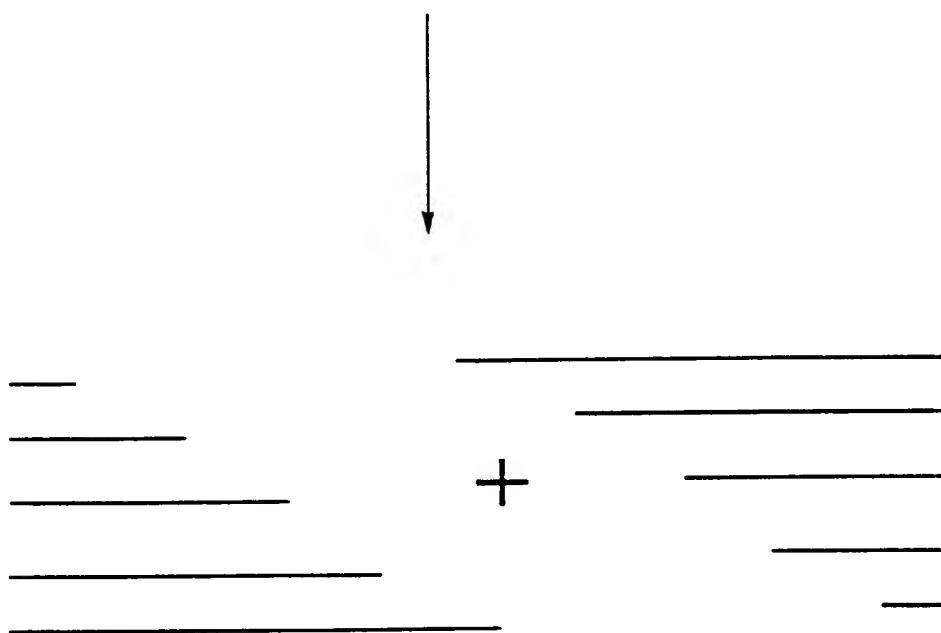
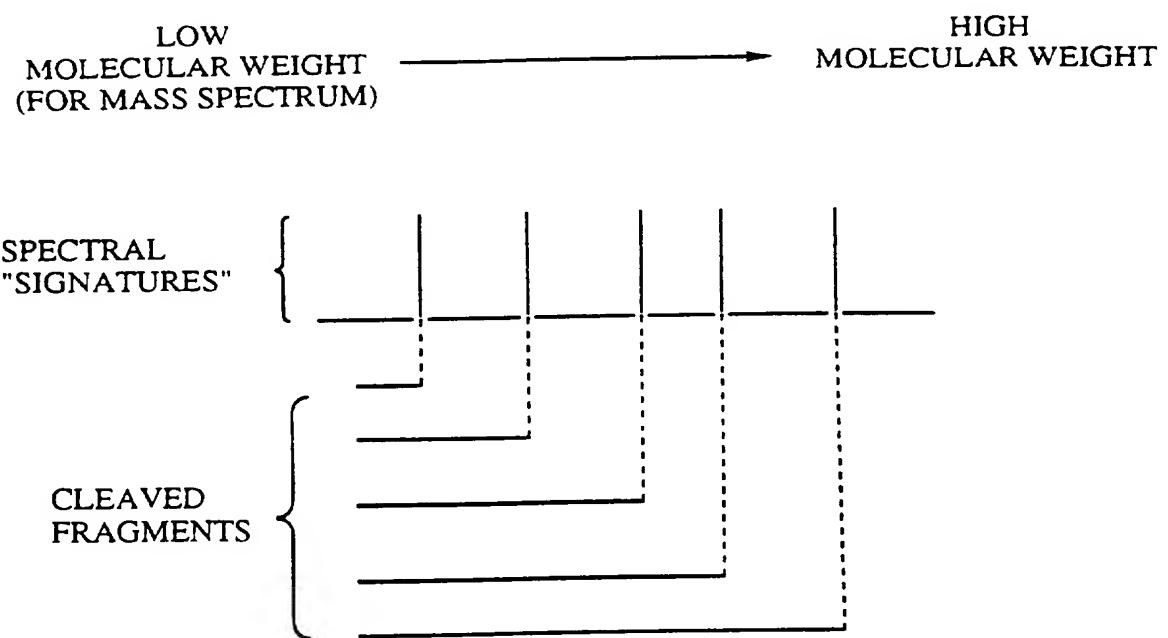


FIGURE 7

8/33

**FIGURE 8**

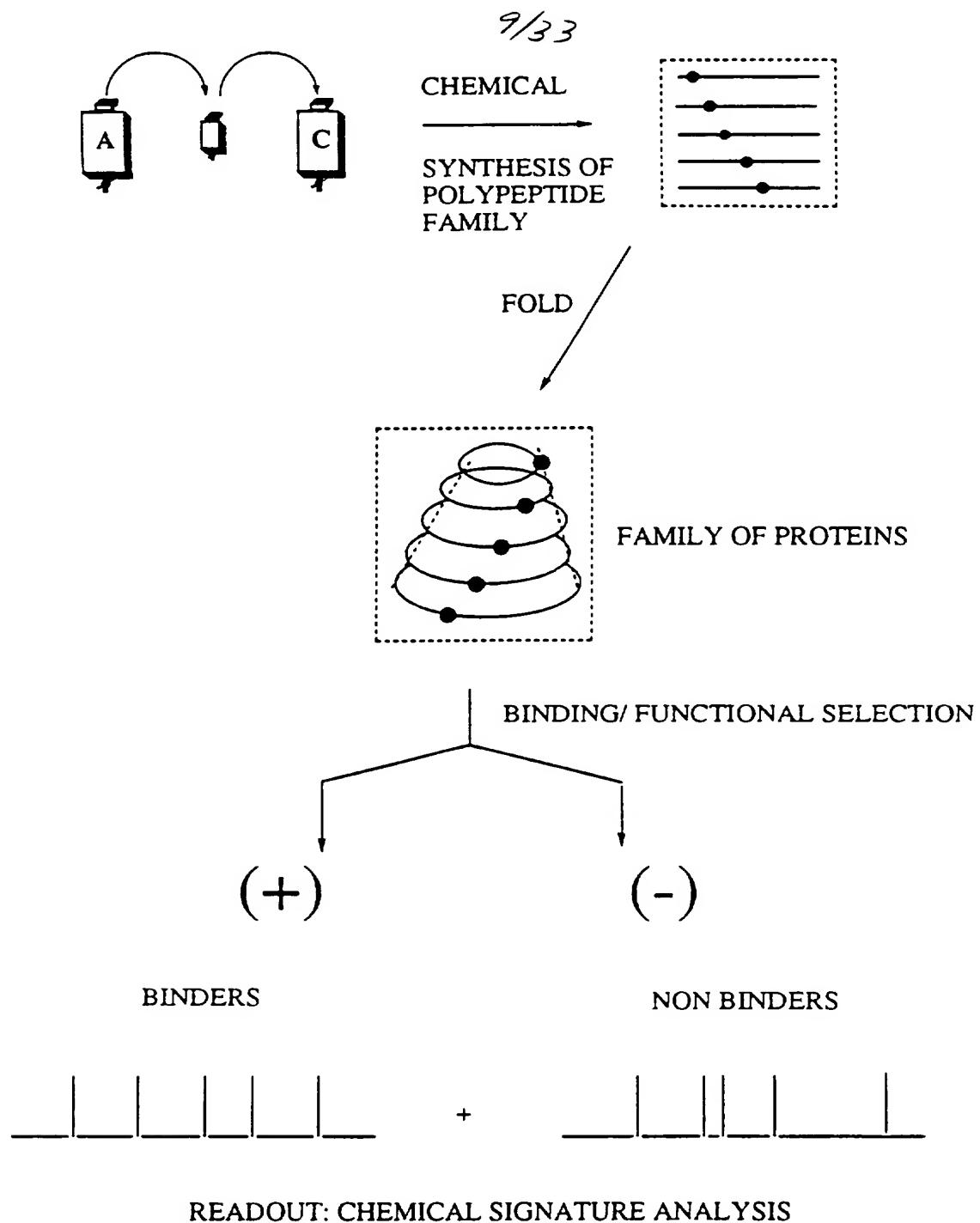


FIGURE 9

10/33

DESIGN

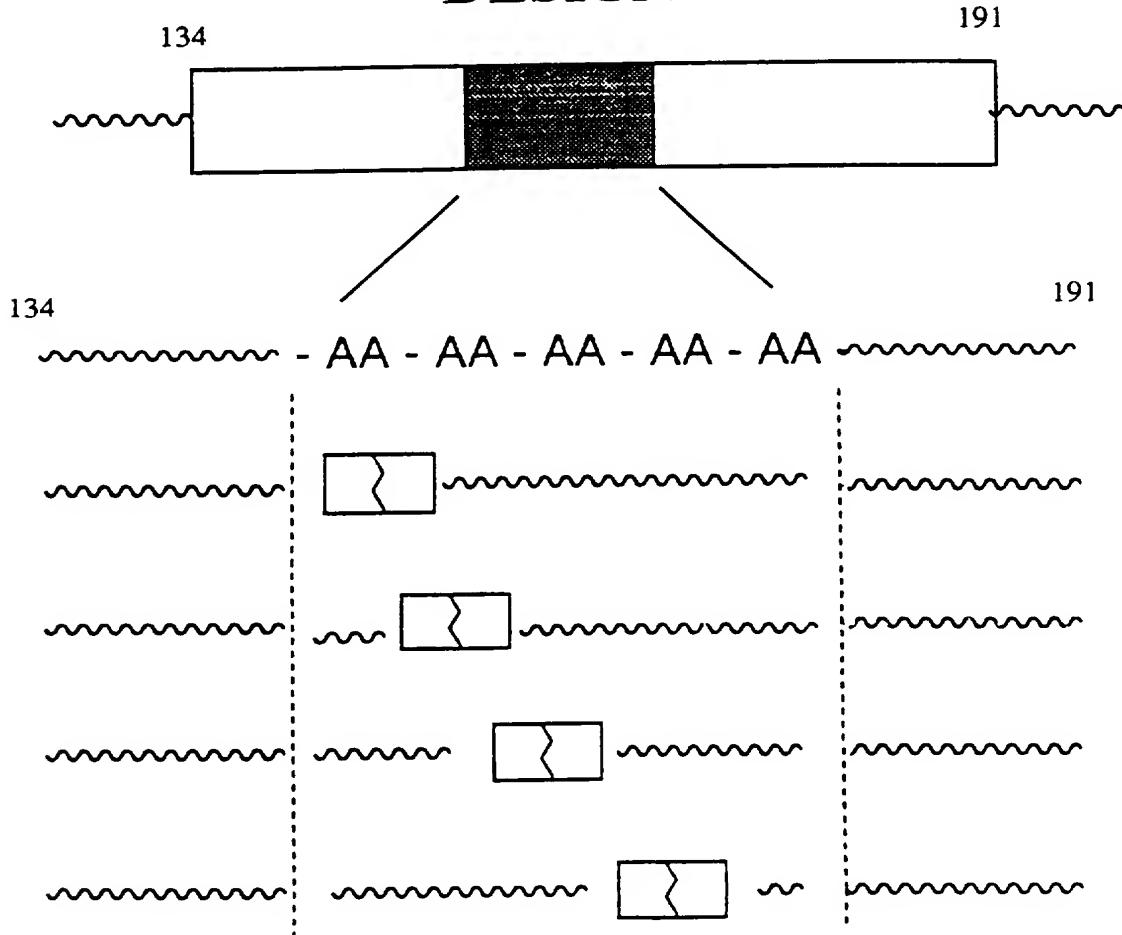


FIGURE 10

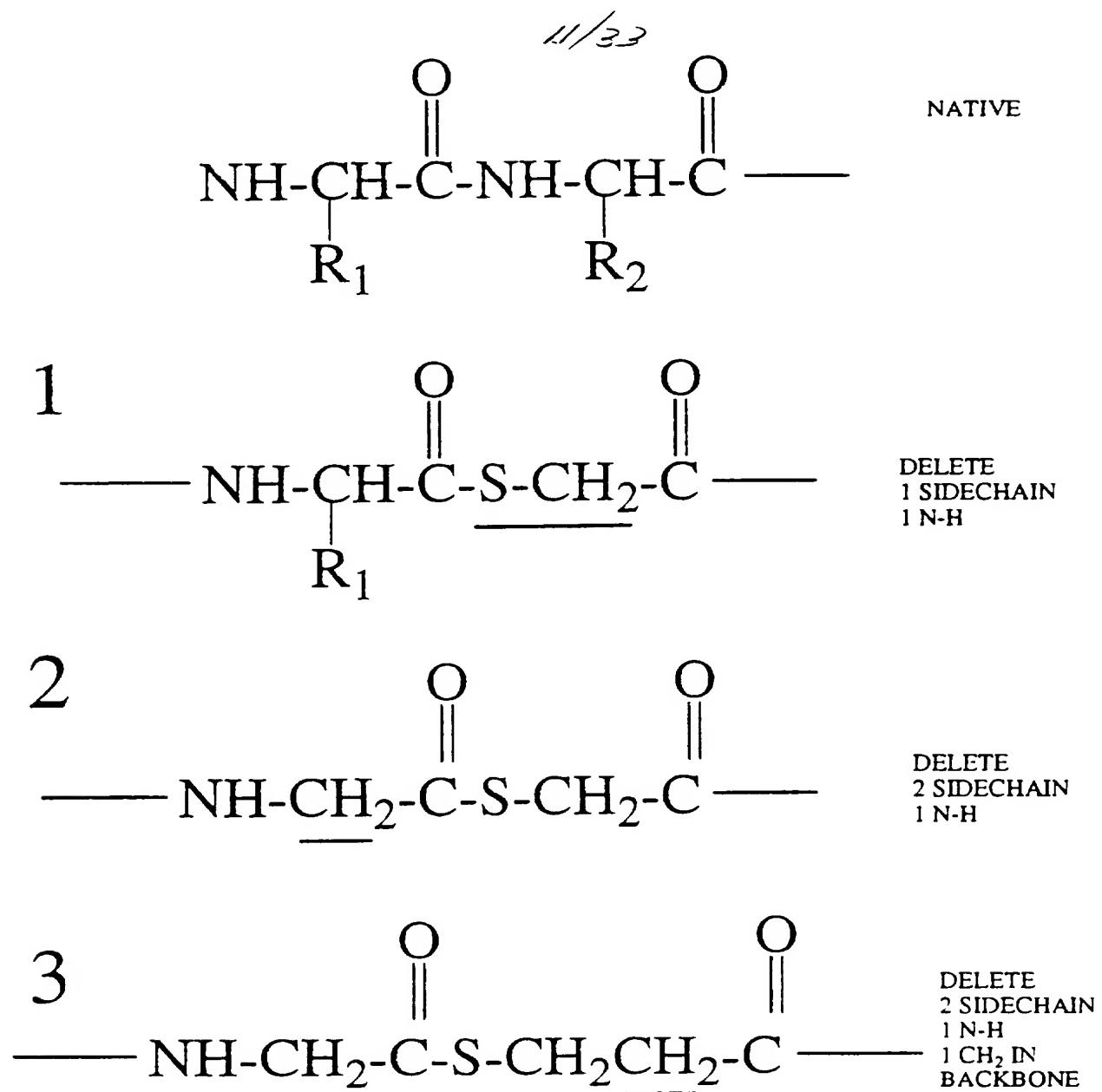


FIGURE 11

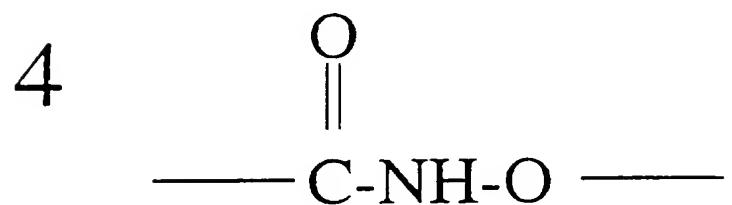
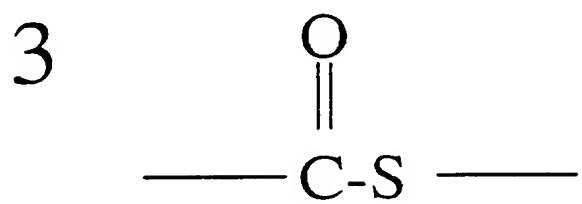
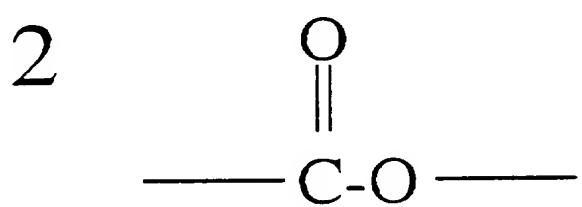
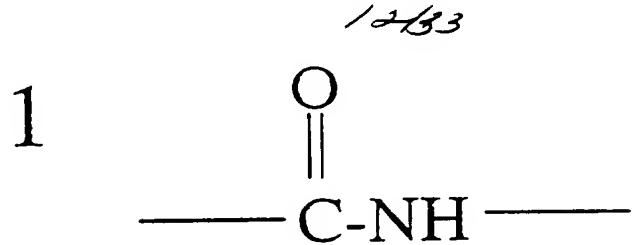


FIGURE 12

13/33

ITERATION:

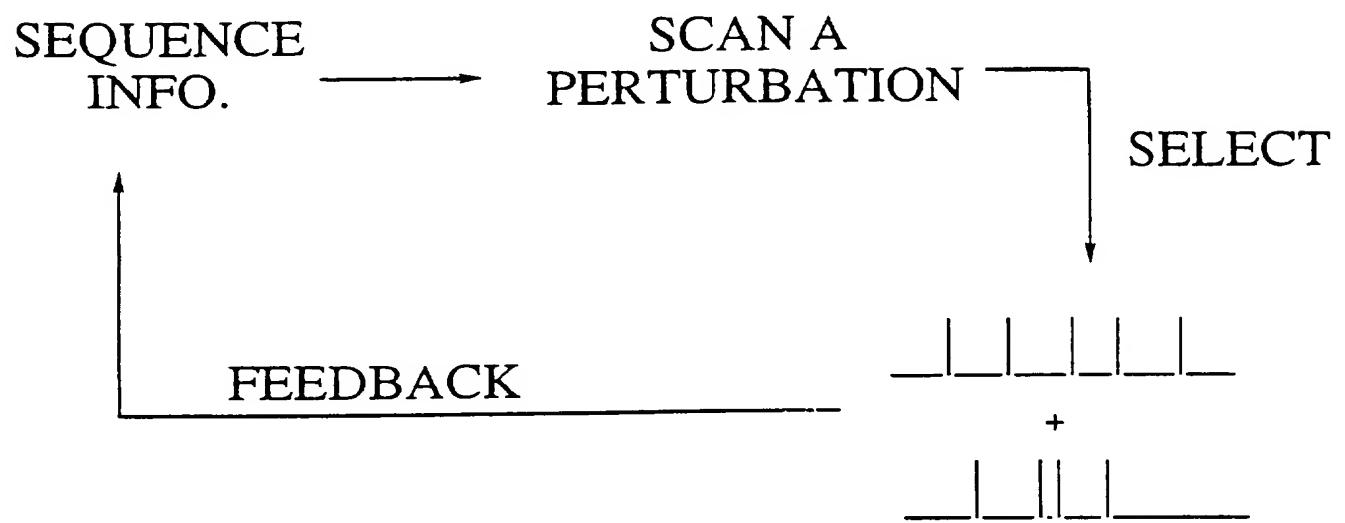


FIGURE 13

14/33

22

~~~~~ -K-  $\triangleright \Sigma$  -I-L-R-I-R-D-K-P ~~~~~

~~~~~ -K-G-  $\triangleright \Sigma$  -L-R-I-R-D-K-P ~~~~~

~~~~~ -K-G-D-  $\triangleright \Sigma$  R-I-R-D-K-P ~~~~~

~~~~~ -K-G-D-I  $\triangleright \Sigma$  -I-R-D-K-P ~~~~~

~~~~~ -K-G-D-I-L  $\triangleright \Sigma$  R-D-K-P ~~~~~

~~~~~ -K-G-D-I-L-R  $\triangleright \Sigma$  D-K-P ~~~~~

~~~~~ -K-G-D-I-L-R-I  $\triangleright \Sigma$  -K-P- ~~~~~

~~~~~ K-G-D-I-L-R-I-R  $\triangleright \Sigma$  -P ~~~~~

~~~~~ -K-G-D-I-L-R-I-R-D  $\triangleright \Sigma$  ~~~~~

FIGURE 14

15/33

**A.**Parent peptide sequence: PFKK-[**GDILRIRDKP**]-EEAcpRLKLKAR

|                                                                                                       |                                                                                                                     |
|-------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------|
| Target<br>Peptide<br>Array                                                                            | PFKK  I L R I R D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G  L R I R D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D  R I R D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I  I R D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I L  R D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I L R  D K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I L R I  K P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I L R I R  P EEAcpRLKLKAR |
|                                                                                                       | PFKK G D I L R I R D  EEAcpRLKLKAR |
|  = Gly-S $\beta$ Ala |                                                                                                                     |

**B.**

| Cycle #                | A to 1       | Neutralize 1 | Couple AA <sub>n</sub> A, B | +BocGly-S $\beta$ Ala to 1 | DMF Wash A, B, 1' | 1' to B | +TFA A, B | 1 to 1' | DMF Wash A, B |
|------------------------|--------------|--------------|-----------------------------|----------------------------|-------------------|---------|-----------|---------|---------------|
| 1) Pro <sup>165</sup>  | 10mL<br>-1mL | ✓            | ✓                           | ✓                          | ✓                 |         | ✓         | ✓       | ✓             |
| 2) Lys <sup>164</sup>  | 9mL<br>-1mL  | ✓            | ✓✓                          | ✓                          | ✓                 | ✓       | ✓✓        | ✓       | ✓✓            |
| 3) Asp <sup>163</sup>  | 8mL<br>-1mL  | ✓            | ✓✓                          | ✓                          | ✓✓✓               | ✓       | ✓✓        | ✓       | ✓✓            |
| etc.                   | ⋮            | ⋮            | ⋮                           | ⋮                          | ⋮                 | ⋮       | ⋮         | ⋮       | ⋮             |
| 9) Asp <sup>157</sup>  | 2mL<br>-1mL  | ✓            | ✓✓                          | ✓                          | ✓✓✓               | ✓       | ✓✓        | ✓       | ✓✓            |
| 10) Gly <sup>156</sup> | ⋮            | ⋮            | ✓✓                          | ⋮                          | ✓✓✓               | ✓       | ✓✓        | ⋮       | ✓✓            |

FIGURE 15

16/33

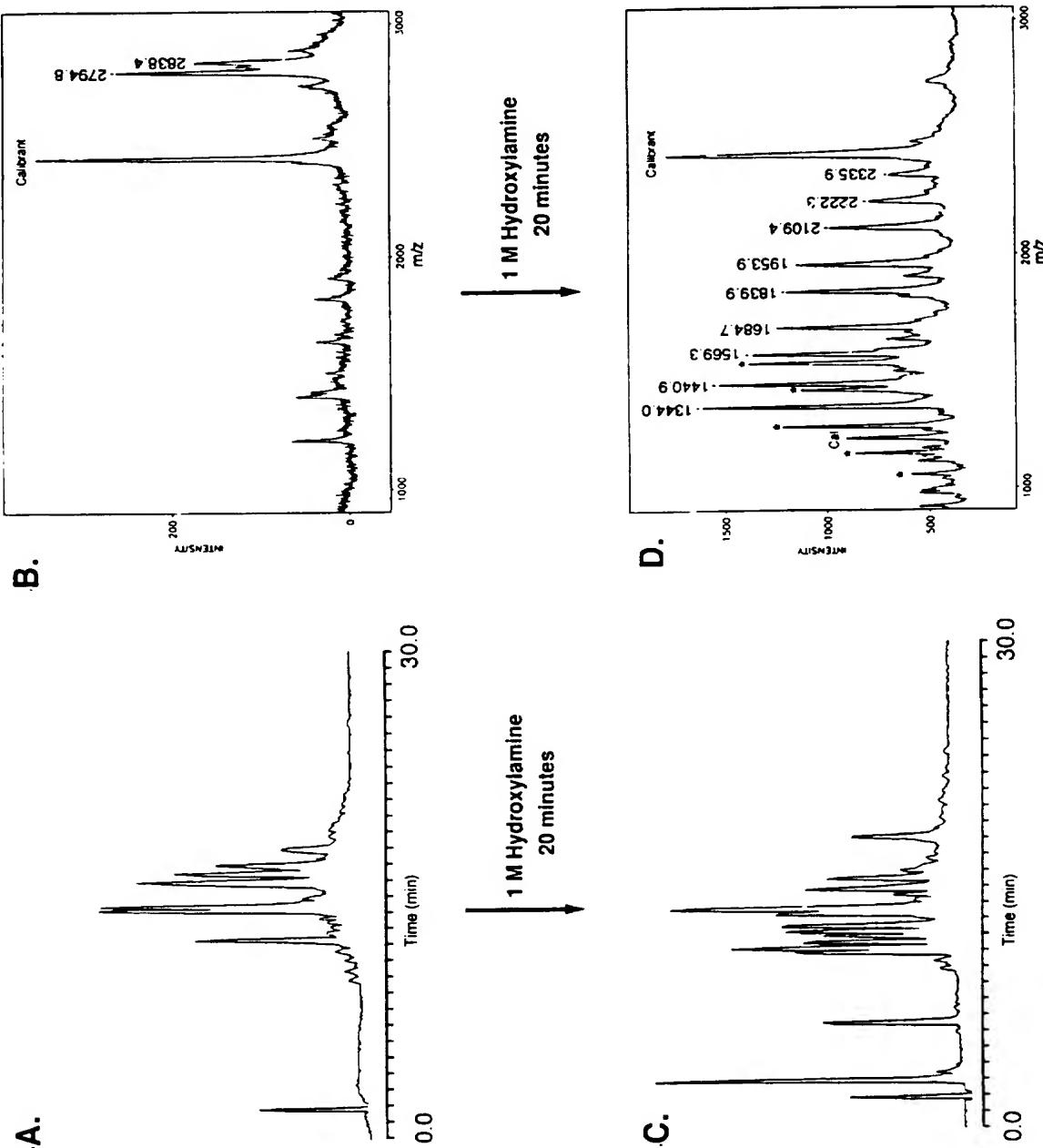


FIGURE 16

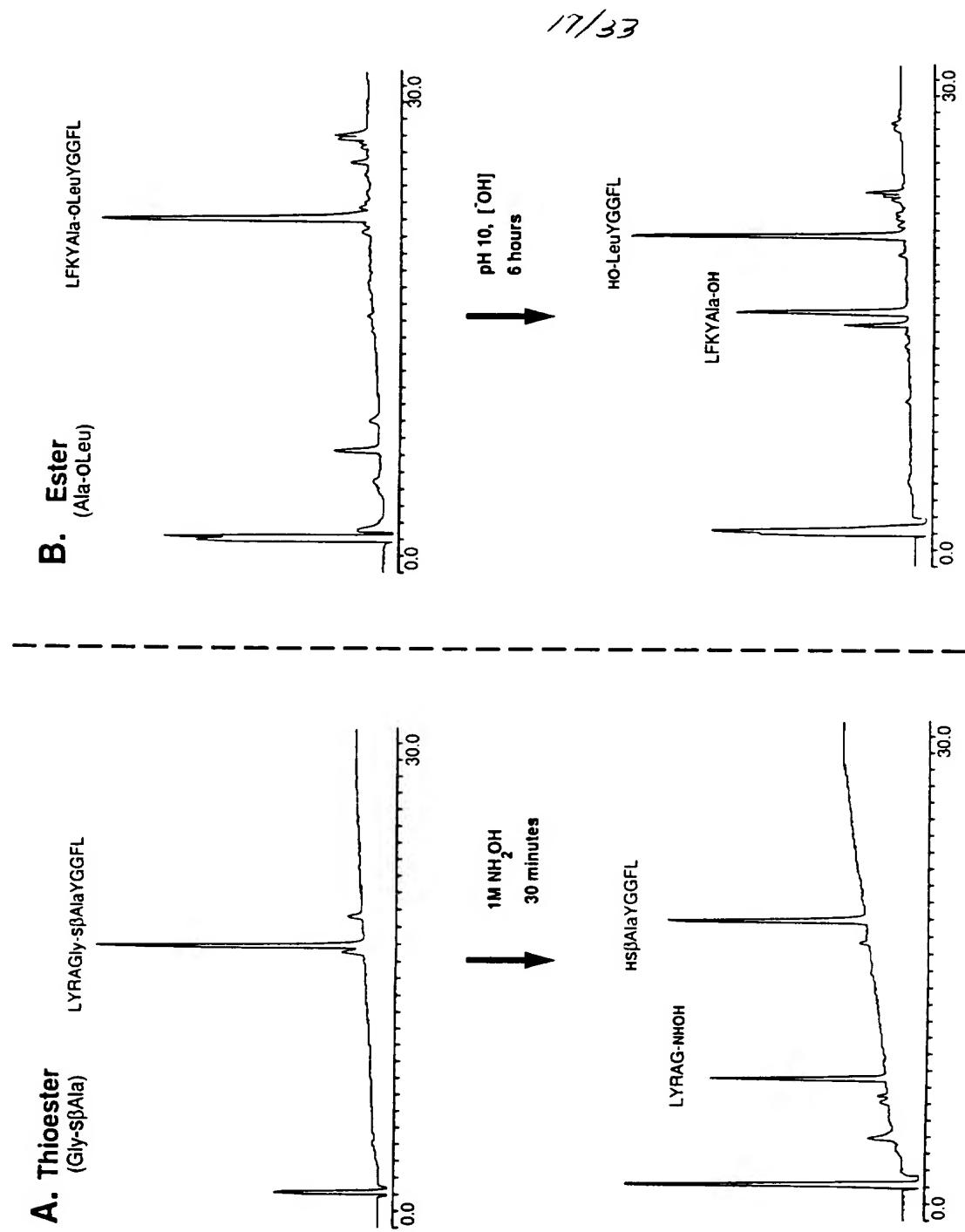
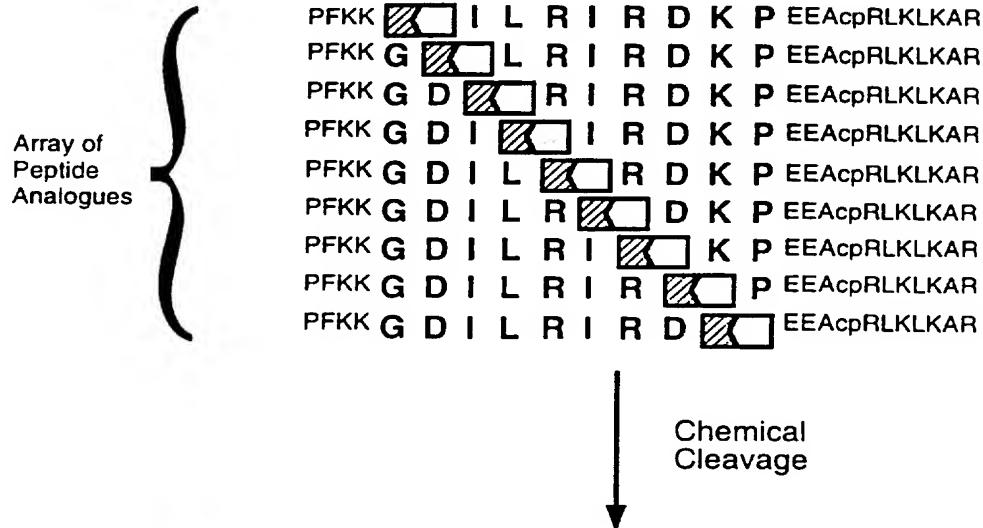


FIGURE 17

18/33

A.



Chemical Cleavage

B.

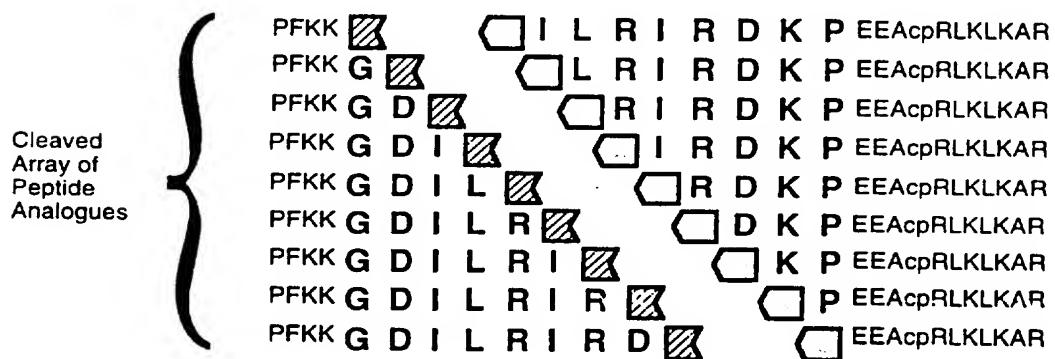


FIGURE 18

19/33

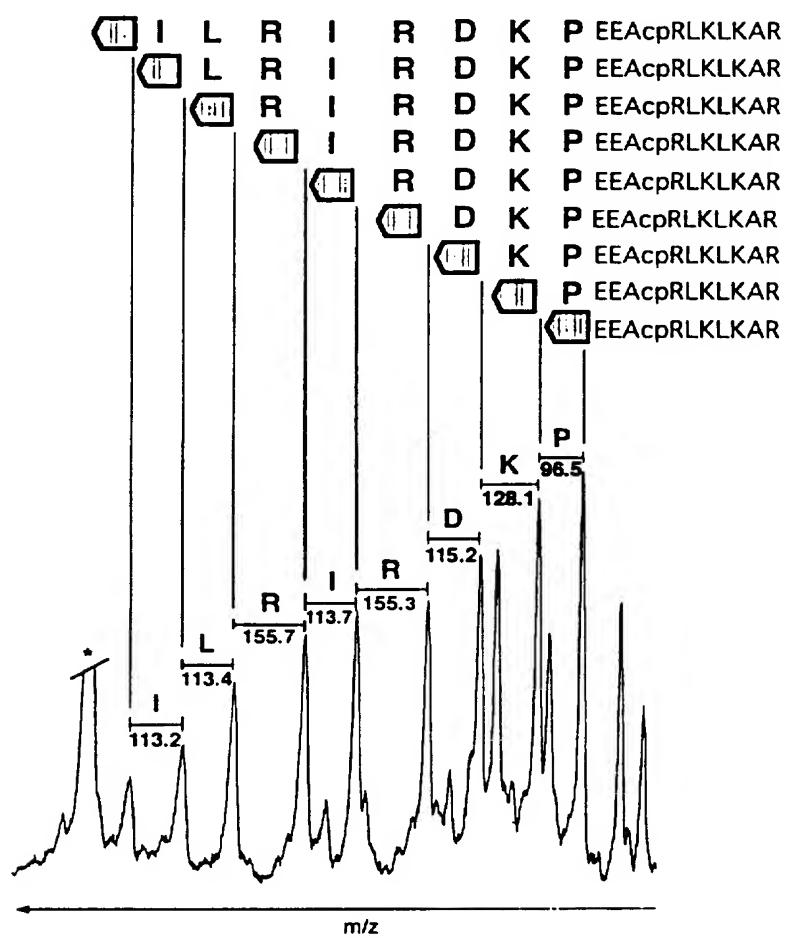


FIGURE 19

## Protein Signature Analysis

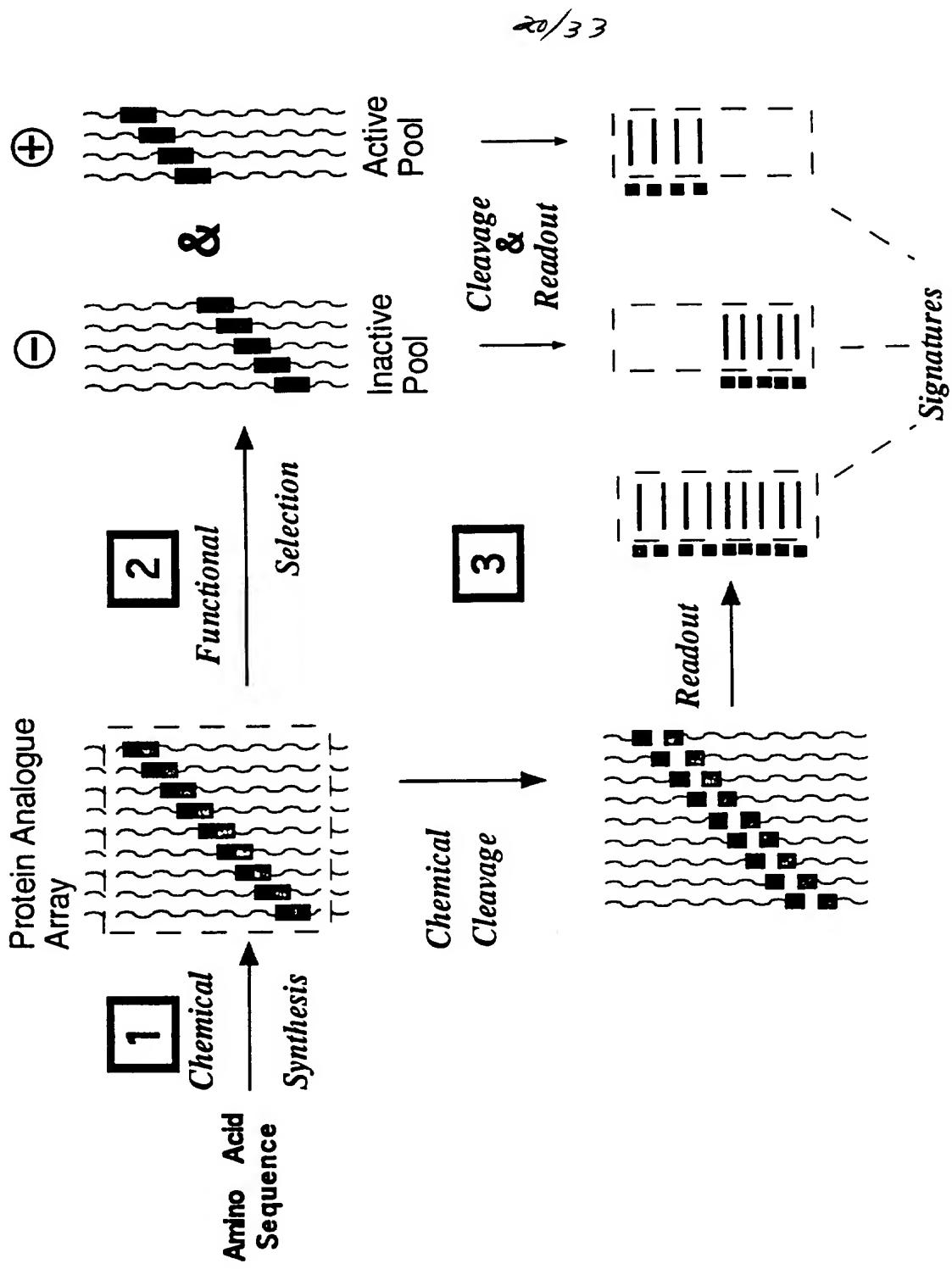


FIGURE 20

21/33

## Protein Analogue Array

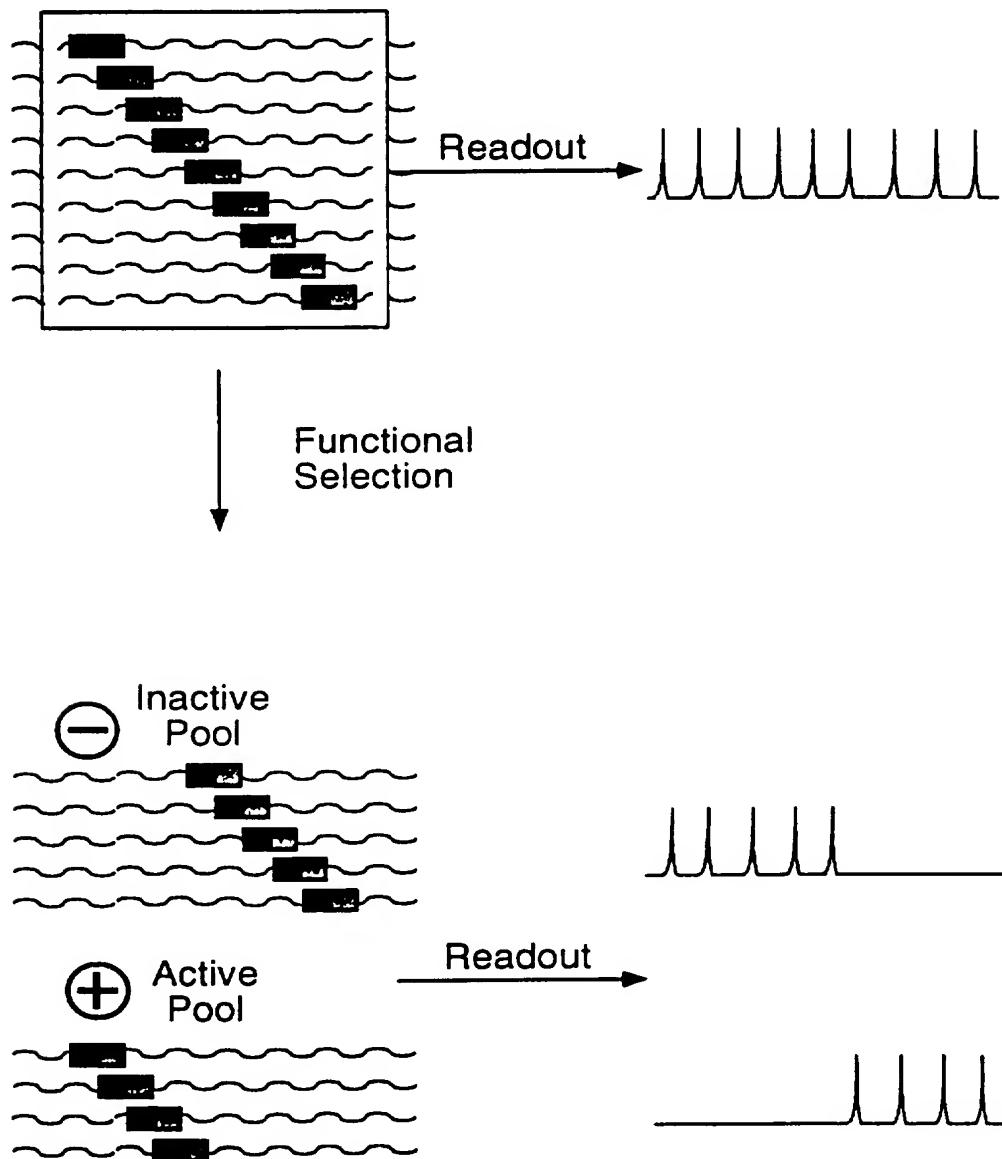


FIGURE 21

22/33

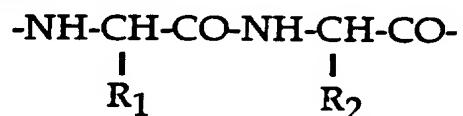
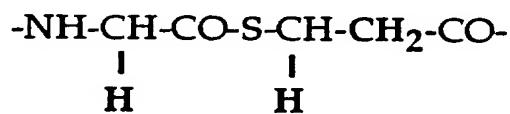
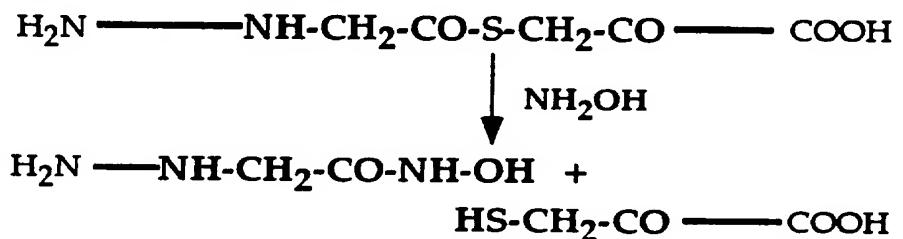
**A.**Native Dipeptide:Gly-S-Gly:Gly-S- $\beta$ Ala:**B.**Chemical Cleavage:

FIGURE 22

23/33

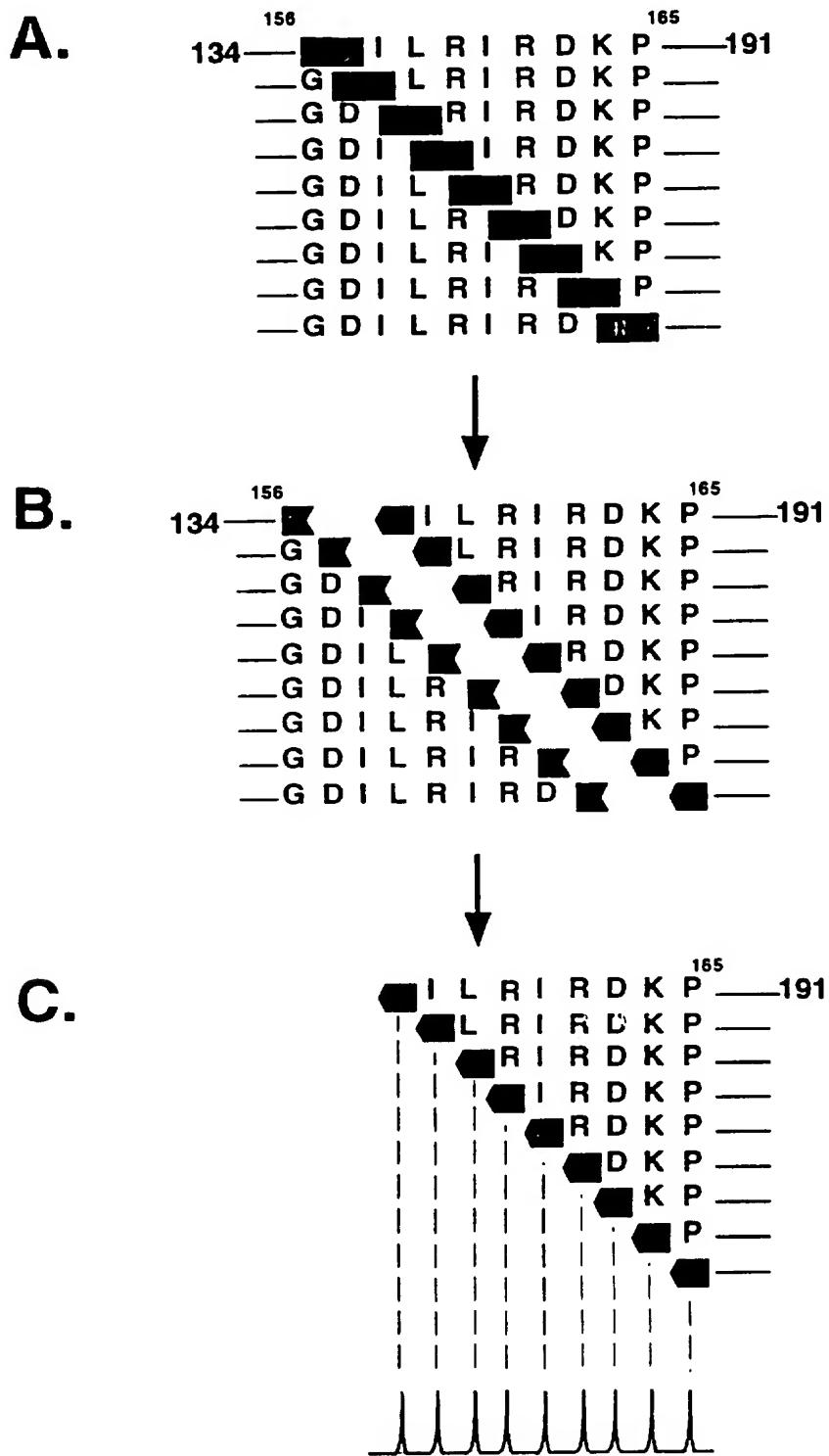


FIGURE 23

24/33

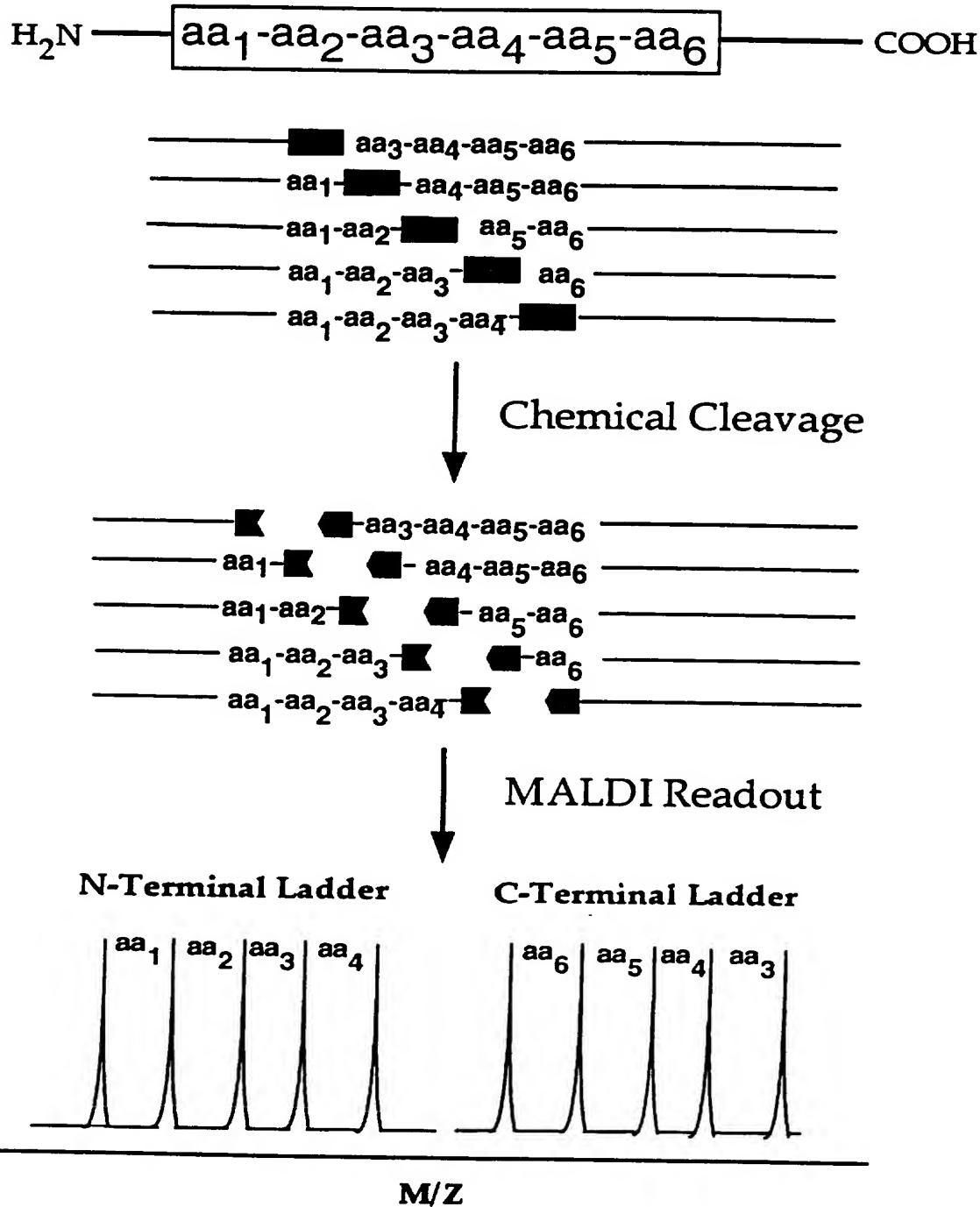
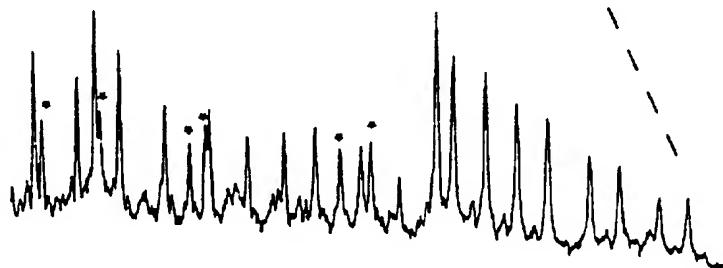


FIGURE 24

25/33

A. 134 146 165 191  
AEYVRALFDNGNDEEDLPFKKGDILRIRDKPEEQWWNAEDSEGKRGMPVVPYVEKVR

B.

C. -Gly-[COS]- $\beta$ Ala-

replaces:

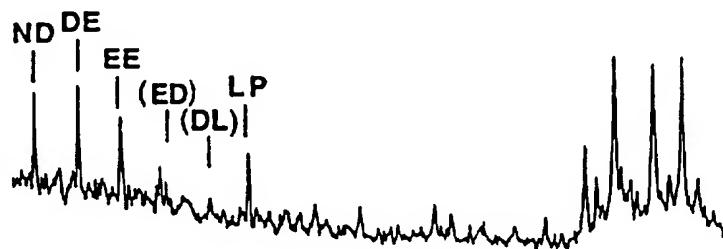


FIGURE 25

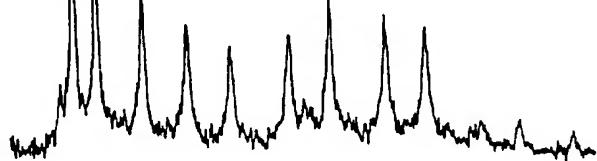
26/33

134 156 165 191  
AEYVRALFDFNGNDEEDLPFKK **GDILRIRDK** PEEQWWNAEDSECKRGMIPVVPYVEKYL

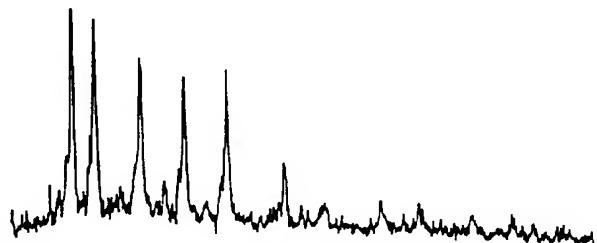
*analogue  
unit  
replaces:*

GD  
| DI IL  
| LR RI IR RD DK KP

A. (control)



B. (wash)



C. (elution)

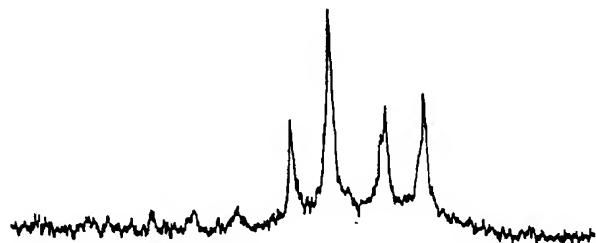
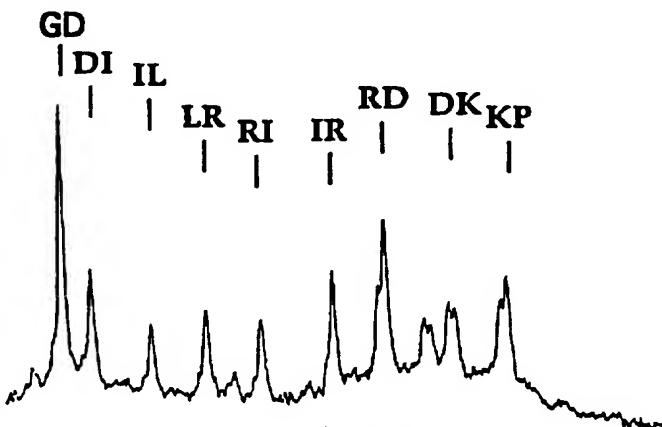


FIGURE 26

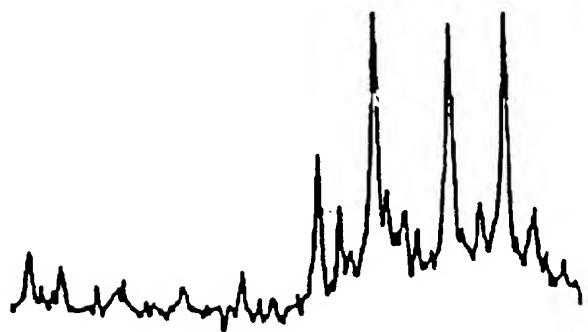
27/33

134                    156                    165                    191  
 AEYVRALFDFNGNDEEDLPFKK**GDILRIRDKP**EEQWWNAEDSEGKRGMPV<sup>P</sup>YVEKYR

*analogue  
unit  
replaces:*



A .-Gly-[COS]-Gly-  
(Binding pool)



B .-Gly-[COS]- $\beta$ Ala-  
(Binding pool)

FIGURE 27

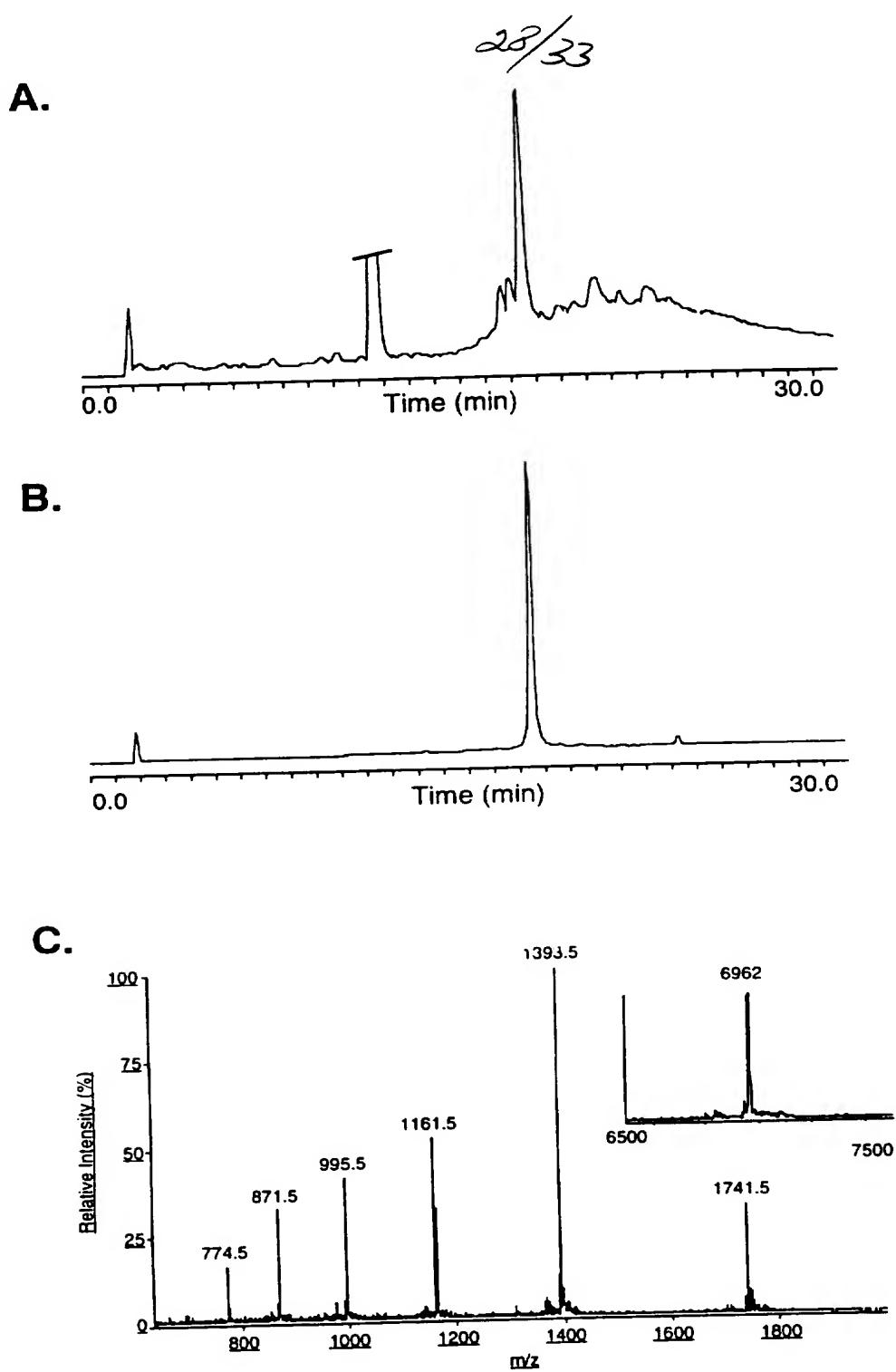


FIGURE 28

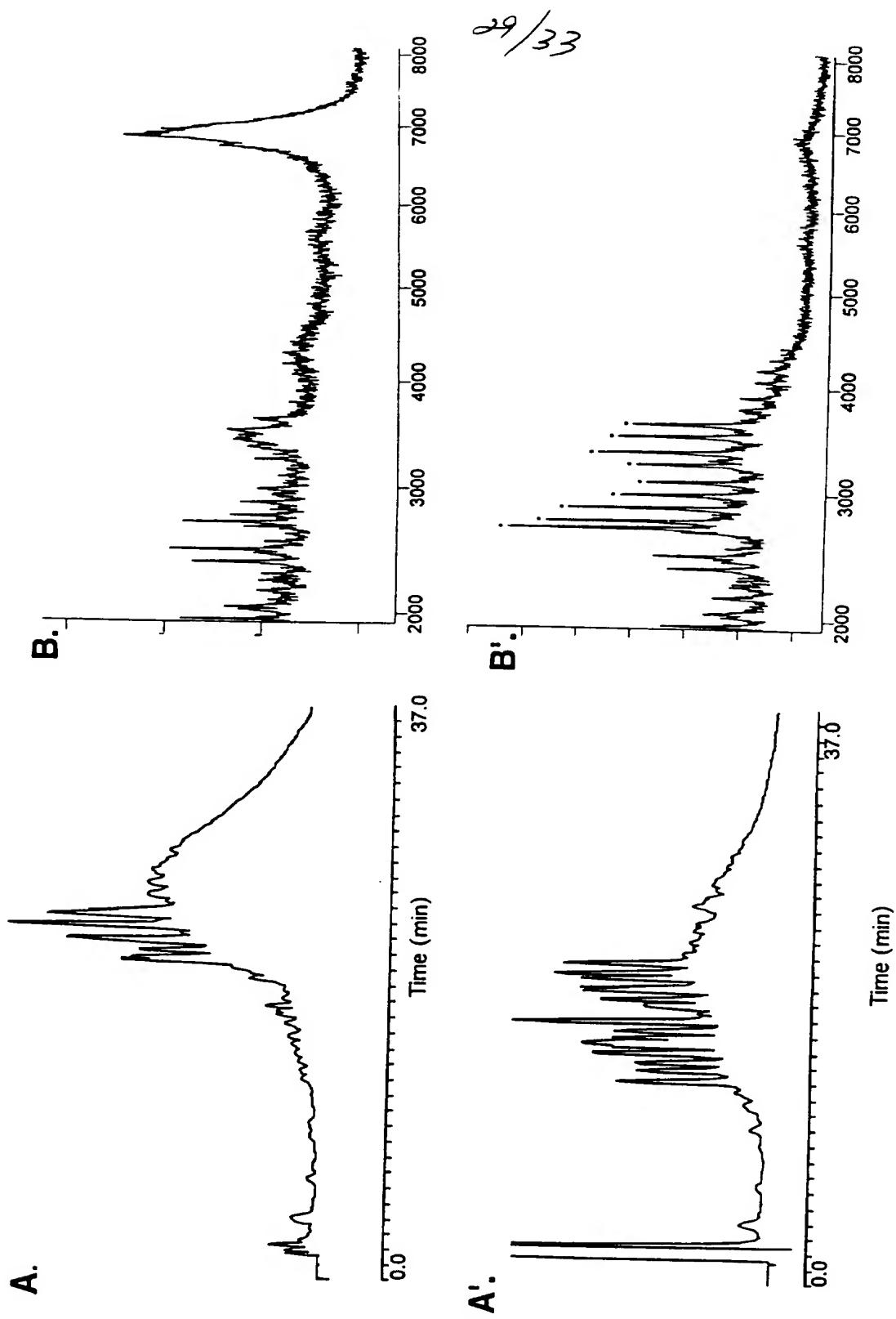


FIGURE 29

30/33

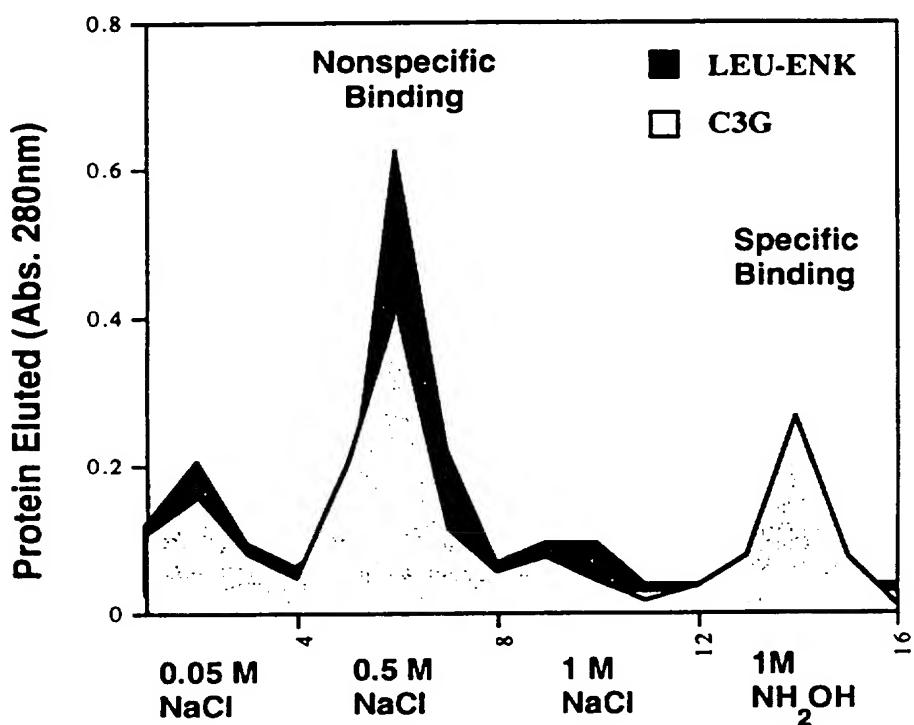


FIGURE 30

31/33

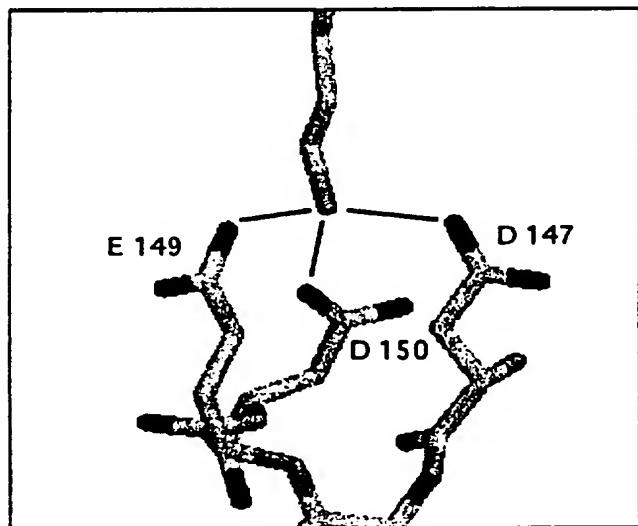


FIGURE 31

32/33

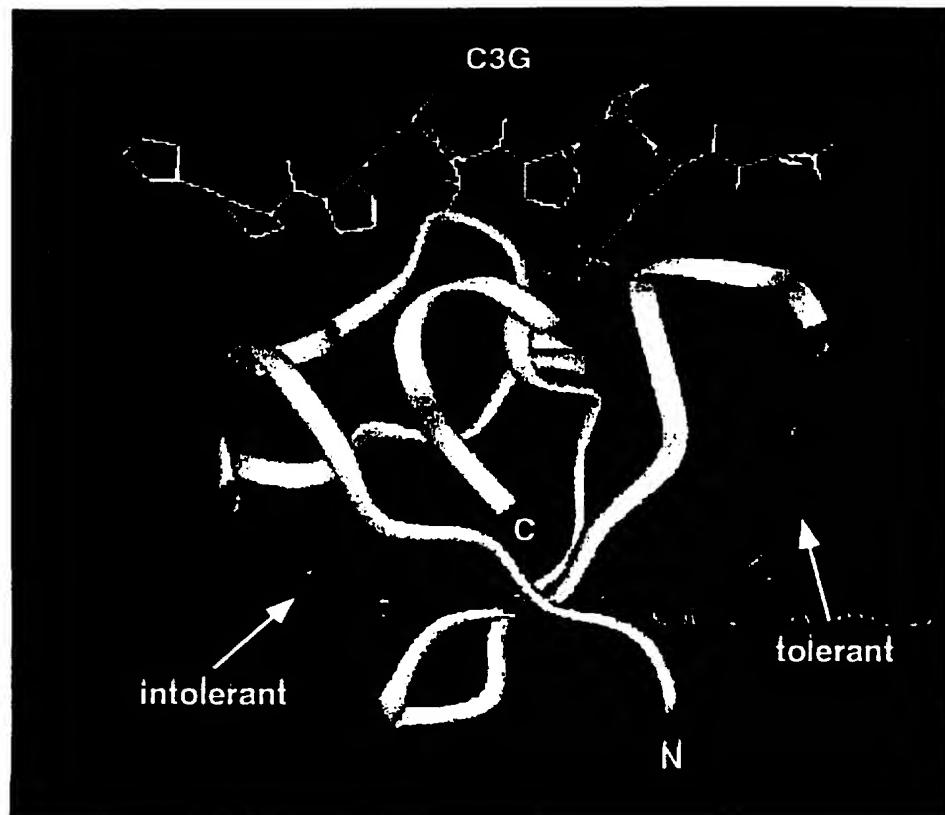


FIGURE 32

33/33

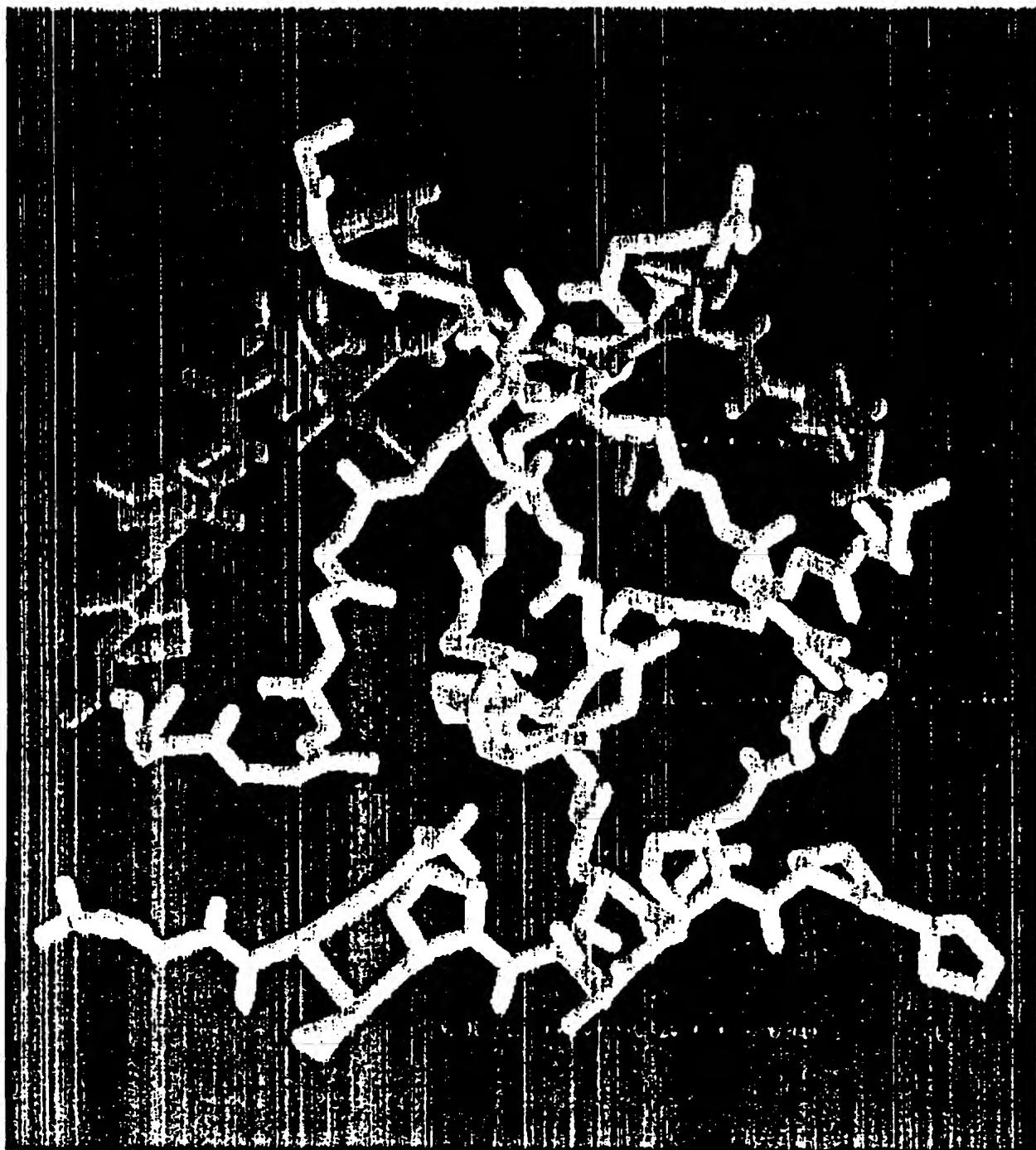


FIGURE 33

## INTERNATIONAL SEARCH REPORT

Intern'l Application No  
PCT/US 96/15516A. CLASSIFICATION OF SUBJECT MATTER  
IPC 6 C07K1/04

According to International Patent Classification (IPC) or to both national classification and IPC

## B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)  
IPC 6 C07K

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)

## C. DOCUMENTS CONSIDERED TO BE RELEVANT

| Category | Citation of document, with indication, where appropriate, of the relevant passages       | Relevant to claim No. |
|----------|------------------------------------------------------------------------------------------|-----------------------|
| A        | WO 95 16209 A (CIBA-GEIGY) 15 June 1995<br>see the whole document<br>---                 | 1-16                  |
| A        | WO 95 11998 A (UNITED BIOMEDICAL INC.) 4<br>May 1995<br>see the whole document<br>---    | 1-16                  |
| T        | WO 96 22067 A (UNITED BIOMEDICAL, INC.) 25<br>July 1996<br>see the whole document<br>--- | 1                     |
|          |                                                                                          | -/-                   |



Further documents are listed in the continuation of box C.



Patent family members are listed in annex.

## \* Special categories of cited documents :

- \*A\* document defining the general state of the art which is not considered to be of particular relevance
- \*E\* earlier document but published on or after the international filing date
- \*L\* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
- \*O\* document referring to an oral disclosure, use, exhibition or other means
- \*P\* document published prior to the international filing date but later than the priority date claimed

\*T\* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

\*X\* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

\*Y\* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

\*&\* document member of the same patent family

1

Date of the actual completion of the international search

11 February 1997

Date of mailing of the international search report

26.02.97

Name and mailing address of the ISA

European Patent Office, P.B. 5818 Patentlaan 2  
NL - 2280 HV Rijswijk  
Tel. (+ 31-70) 340-2040, Tx. 31 651 epo nl,  
Fax (+ 31-70) 340-3016

Authorized officer

Masturzo, P

# INTERNATIONAL SEARCH REPORT

Internal Application No  
PCT/US 96/15516

| C.(Continuation) DOCUMENTS CONSIDERED TO BE RELEVANT |                                                                                                                                                                                                                                                                                                                                                                               |                       |
|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| Category *                                           | Citation of document, with indication, where appropriate, of the relevant passages                                                                                                                                                                                                                                                                                            | Relevant to claim No. |
| T                                                    | <p>BIOLOGICAL ABSTRACTS, vol. 103, no. 1,<br/>1997<br/>Philadelphia, PA, US;<br/>abstract no. 1173,<br/>T W MUIR ET AL.: "Probing the chemical<br/>basis of binding activity in an SH3 domain<br/>by protein signature analysis"<br/>XP002025005<br/>see abstract<br/>&amp; CHEMISTRY AND BIOLOGY,<br/>vol. 3, no. 10, October 1996, LONDON,<br/>pages 817-825,<br/>-----</p> | 1-16                  |
| 1                                                    |                                                                                                                                                                                                                                                                                                                                                                               |                       |

# INTERNATIONAL SEARCH REPORT

Information on patent family members

Intern'l Application No

PCT/US 96/15516

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s) |         | Publication<br>date |
|-------------------------------------------|---------------------|----------------------------|---------|---------------------|
| WO-A-9516209                              | 15-06-95            | AU-A-                      | 1068295 | 27-06-95            |
|                                           |                     | CA-A-                      | 2175078 | 15-06-95            |
|                                           |                     | EP-A-                      | 0733213 | 25-09-96            |
|                                           |                     | FI-A-                      | 962371  | 07-06-96            |
|                                           |                     | NO-A-                      | 962186  | 29-05-96            |
| -----                                     | -----               | -----                      | -----   | -----               |
| WO-A-9511998                              | 04-05-95            | AU-A-                      | 8091694 | 22-05-95            |
|                                           |                     | EP-A-                      | 0725838 | 14-08-96            |
| -----                                     | -----               | -----                      | -----   | -----               |
| WO-A-9622067                              | 25-07-96            | AU-A-                      | 5849796 | 07-08-96            |
| -----                                     | -----               | -----                      | -----   | -----               |